Impact of STARD on reporting quality of diagnostic accuracy studies in a top Indian Medical Journal: A retrospective survey

Rajashree Yellur [1] , **Shabbeer Hassan** Corresp. [2]

[1] Department of Statistics, Manipal University, Manipal Academy of Higher Education, Manipal, India
[2] Institute for Molecular Medicine, University of Helsinki,, University of Helsinki, Helsinki, Finland

Corresponding Author: Shabbeer Hassan
Email address: shabbeer.hassan@helsinki.fi

Improper reporting of diagnostic studies leads to an incorrect assessment of their clinical performance. STARD (Standards for Reporting of Diagnostic Accuracy Studies) checklist was launched in 2003 with the intention of improving reporting quality in diagnostic accuracy studies. The main aim of this study was to check the extent to which published diagnostic accuracy studies follow the 28-item STARD checklist. We conducted a literature survey of diagnostic studies published in Indian Journal of Medical Research (IJMR) between the years 1995-2013 for the evaluating their reporting quality by checking their adherence to STARD. Relevant studies (N=76) were retrieved from IJMR website and data extraction was performed by two authors simultaneously. A simple pre-post analysis found that there was no overall change in the reporting quality before and after STARD was released. Though some STARD items like description of participant sampling ($\chi^2 = 5.712$, $p = 0.0169$), clinical applicability of study findings ($\chi^2 = 9.704$, $p = 0.0018$) had a significant increase in post-STARD period. To take into account any underlying trend we conducted an interrupted time-series was done. We found a significant increase in the reporting quality after publication of STARD ($\beta_3 = 0.215 \pm 0.068$, $p = 0.034$). The overall reporting quality of diagnostic accuracy studies have improved since the introduction of STARD, however, error/defects in many sections remain as before.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3413v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

IMPACT OF STARD ON REPORTING QUALITY OF DIAGNOSTIC ACCURACY STUDIES IN A TOP INDIAN MEDICAL JOURNAL: A RETROSPECTIVE STUDY

Rajashree Yellur, Shabbeer Hassan

Rajashree Yellur, MSc Biostatistics, Department of Statistics, Level 6, Health Science Library Building, Manipal University, Manipal 576104.

Email: rajashreeyellur@gmail.com

Shabbeer Hassan, FIMM-EMBL PhD student, Institute for Molecular Medicine (FIMM), University of Helsinki, Finland-00014.

Email: shabbeer.hassan@helsinki.fi

CORRESPONDING AUTHOR

Shabbeer Hassan,

FIMM-EMBL PhD student, Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland-00014

Email: shabbeer.hassan@helsinki.fi

Phone: +358-468437288

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3413v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

ABSTRACT

Improper reporting of diagnostic studies leads to an incorrect assessment of their clinical performance. STARD (Standards for Reporting of Diagnostic Accuracy Studies) checklist was launched in 2003 with the intention of improving reporting quality in diagnostic accuracy studies. The main aim of this study was to check the extent to which published diagnostic accuracy studies follow the 28-item STARD checklist. We conducted a literature survey of diagnostic studies published in Indian Journal of Medical Research (IJMR) between the years 1995-2013 for the evaluating their reporting quality by checking their adherence to STARD. Relevant studies (N=76) were retrieved from IJMR website and data extraction was performed by two authors simultaneously. A simple pre-post analysis found that there was no overall change in the reporting quality before and after STARD was released. Though some STARD items like description of participant sampling ($\chi^2 = 5.712$, $p = 0.0169$), clinical applicability of study findings ($\chi^2 = 9.704$, $p = 0.0018$) had a significant increase in post-STARD period. To take into account any underlying trend we conducted an interrupted time-series was done. We found a significant increase in the reporting quality after publication of STARD ($\beta_3 = 0.215 \pm 0.068$, $p = 0.034$). The overall reporting quality of diagnostic accuracy studies have improved since the introduction of STARD, however, error/defects in many sections remain as before.

INTRODUCTION

Diagnostic studies are conducted to evaluate how efficacious a given test is in reference to a given disorder. A better nomenclature for them however is diagnostic accuracy studies. Here, accuracy refers to the rate of agreement between the current test under evaluation, known as index test and a standard test or gold/reference standard. Diagnostic accuracy from such kind of studies are usually reported as: sensitivity, specificity, likelihood ratio, AUC etc. [Griner et al., 1984; Metz 1978; Sackett et al., 1991]

The clinician uses this information to make decisions whether a given diagnostic test is useful for a given disorder or not. Hence, badly conducted or reported diagnostic studies would lead to biased results, which in turn might mislead clinicians endangering patients' lives [Lijmer et al., 1999].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3413v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Several factors are known to affect the internal and external validity of diagnostic accuracy studies. Several reviews [Lijmer et al., 1999; Reid et al., 1995; Plint et al., 2006; Moher et al., 2001; Turner et al., 2012] which looked at the reporting of diagnostic studies found that several major elements like design, conduct or analysis are missing and not reported at all.

The STARD (Standards for Reporting of Diagnostic Accuracy Studies) statement was published in 2003 as a public release in 13 reputed biomedical journals. The primary aim was to combat the growing menace of incomplete reporting and poorly designed diagnostic accuracy studies as reported by some reviews published before STARD checklist was published [Bossuyt, 2008]. The checklist contains 28 items for inclusion by authors which should be then checked by journal reviewers.

Apart from the checklist, STARD prescribes a flowchart similar to the PRISMA statement which describes the flow of participant inclusion/exclusion in the study. Till now, around 200 journals have supported the STARD statement (http://www.stard-statement.org/).

In the past 20 years many reporting guidelines like CONSORT for randomized controlled trials, STROBE for observational studies etc. have been introduced. Since then, many researchers have conducted studies to test the impact of such guidelines on the reporting quality of published studies but results so far have been conflicting at best.

For example, in case of STARD guideline, there have been controversies surrounding its impact as one study saw a minor increase in the reporting quality after STARD [Smidt et al., 2006] whereas another study didn’t find it to be the case [Wilczynski et al., 2008]. We believe this controversy might be due to ignoring the underlying time trend underlying the reporting quality change. To address this issue, we used interrupted time series analysis apart from the normal pre-post statistical test.

We decided to focus on a single medical journal to test the role of STARD in changing if any, the reporting quality of published diagnostic studies. Indian Journal of Medical Research (IJMR) is one of India’s and in fact one of Asia’s best medical journals with more than 100 years of publication history. It has one of the highest impact factors among Indian medical journals (http://www.icmr.nic.in/Publications/IJMR.html). Because of its widespread reputation and readership among clinicians we decided to focus our evaluation of the reporting quality of diagnostic accuracy studies only on IJMR.

METHODS

Search Criteria

To identify all the eligible studies, we conducted a PubMed search of IJMR and manually searched all issues of the journal published during the study period. The keyword used for the search in Pubmed were (("sensitivity AND specificity" OR "specificit* " OR "false negative" OR "accuracy")) AND "Indian Journal of Medical Research"[Journal] with studies restricted/limited to humans and only those studies with abstracts''.

Article Selection

We selected all articles published between January 1999 and December 2013 that were declared as diagnostic studies or used sensitivity or specificity in their preferred mode of analysis. The analysis time period was chosen in such a way that it formed an approximate 10-year window around the release of STARD. We did not select any letters to editors, or review papers. The titles and abstracts were screened by two of us (SH and RY) working independently of each other and resolving disagreements by consensus, which led to the selection of 76 articles. The names and affiliations of the authors and the dates of article acceptance and publication were masked to minimize evaluation bias by the raters.

Data abstraction

We included all 25 items in the STARD checklist along with three additional items from other published checklists to represent the changing demands of a published article. Each article was evaluated based on the 28 items of our checklist (Table 1). Further on, each item in the checklist was evaluated using a three-point rating scale: 1- criteria met, 2 - criteria not met, and 3 - cannot determine or not relevant. All problems were reviewed by the authors (SH and RY) within themselves and external faculty from Department of Statistics, Manipal University served as the final adjudicator. Data was collected using a user-friendly form with EpiData version 3.1.

Outcome measure

The primary outcome is a composite score obtained from our checklist defined as the number of the 28 items properly reported divided by the total number of applicable items. Here total applicable items was found out by subtracting total to the number of non-applicable items for each article. The score was then expressed in form of a percentage. This study did not require approval by an ethics committee, since it concerned research publications and not individuals. The inter-rater agreement for all the information coded from the articles was examined using the intraclass correlation coefficient (ICC) [Shrout et al., 1979]. The ICC values for the 28 items related to the diagnostic studies adherence to STARD ranged from 0.83 to 0.989.

Data analysis

All the quantitative variables are summarized here as mean (standard deviation) and qualitative variables as number (percentage). We used a paired t-test to determine whether there was a change in the outcome before (1995-2002) and after (2004-2013) STARD publication and Mc-Nemar test for testing change in outcome of certain items within our checklist.

Interrupted time series analysis: To address the underneath time trend of the reporting quality after STARD publication we also conducted an interrupted time-series analysis using a segmented regression model. The main question was to determine whether STARD had any impact on the mean score after its publication [Eccles et al., 2003; Ramsay et al., 2003; Wagner et al., 2002]. We considered two periods, pre (1993-2002) and post- STARD period (2004-2013). In the model, dependent variable was the checklist score mean and the independent variable was year considered.

The segmented regression model as in [Cochrane ITS study, 2009]:

$$\text{Mean Score} = \text{Constant} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Here the coefficient for ‘X_1’ (β_1) gives the slope of the regression line pre- STARD, coefficient for ‘X_2’ (β_2) is the change in intercept and coefficient for ‘X_3’ (β_3) provides the change in slope pre-and post STARD.

Therefore, pre STARD: Outcome = constant + β_1*time and post STARD: Outcome = Constant + $\beta_1 X_1 + \beta_2 + \beta_3 X_3 = (\text{constant} + \beta_2) + (\beta_1 + \beta_3)$*time (as X_1 and X_3 remain the same post STARD). Therefore, the difference in constant (intercept) pre-and post STARD is β_2 and difference in slope is β_3. The level and trend of pre- STARD segment (1995 -2002) served as the control for the post-STARD segment (2002- 2013). We estimated the difference between pre- STARD and post-

STARD slopes and the yearly mean effect after STARD publication. Durbin-Watson test was used to test the residual independence.

Only two-tailed tests were used and p-values less than 0.05 was taken as to be statistically significant. The analysis was conducted using IBM SPSS v16.0 (Armonk, New York, U.S).

RESULTS

Seventy-six (76) articles were downloaded and their data extracted from Indian Journal of Medical Research (IJMR) for the period of 1995-2013. A complete list of the articles used here for analysis is available from the first author. The percentage of articles meeting each of the 28 criteria (Table 1) for the whole timeline (1995-2013) is presented in Table 2.

Descriptives

STARD: Introduction

In the years before the STARD release around 29 %(10/35) articles identified themselves as diagnostic accuracy studies whereas post STARD around 42.5% (17/41) articles identified them so. However, we found this difference to be statistically insignificant (χ^2 = 0.009, p = 0.9243).

Around 42.9 %(15/35) articles had clear aims and stated the research questions clearly but this figure didn’t change post-STARD 44 %(18/41).

STARD: Methods

The method section of the STARD checklist in Table 1 has been divided into various subsections: participants, test methods and statistical methods. There were no changes observed both in pre- and post STARD period in regards to description of study population (χ^2 = 0.567, p = 0.1158), participant recruitment (χ^2 = 0.172, p = 0.6784), adequate sampling (χ^2 = 0.421, p = 0.2447), sample size calculation, description of data collection (χ^2 = 0.386, p = 0.5345), description of reference standard and its underlying rationale (χ^2 = 0.357, p = 0.55) and description of the technical specifications (χ^2 = 0.22, p = 0.0719). In statistical methods, no statistically significant change was observed in reporting of the methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (χ^2 = 0.029, p = 0.1158).

However, significant improvement in mentioning the software used to conduct the analysis was found in the post-STARD period as compared to the pre-STARD period ($\chi^2 = 9.122$, p = 0.0014). Also, in terms of description of participant sampling, post-STARD period saw a significant change ($\chi^2 = 5.712$, p = 0.0169).

STARD: Results

In regards to the description of results in diagnostic studies (Item nos. in Table 1: 17-21 and 23-27) within IJMR no statistical change was observed between pre-and post STARD period. However, there was a significant change (p = 0.0045) in post-STARD period for the reporting of cross tabulation of the results of the index tests by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard. Changes in these key items within STARD between pre-post periods is presented in Figure 1.

STARD: Discussion

A major change in this section has been that post-STARD, increasingly articles have been discussing the clinical applicability of study findings ($\chi^2 = 9.704$, p = 0.0018).

Interrupted Time Series Analyses

The above analyses use scores averaged over the pre-post STARD period which were then compared for any statistically significant changes. As mentioned before, a majority of review literature on various guidelines use such kind of average based statistics. Here, we used an interrupted time-series analyses which can detect whether STARD publication had a significant effect than the underlying trend [17]. Here we considered two periods: pre (1993-2002) and post-STARD (2004-2013) period.

In the pre STARD period, the mean score increased non-significantly (p = 0.124). This trend did not change significantly after publication of the STARD statement until 2010 (p = 0.067 for year 2010). However, from that point of time onwards we see there is a significant change in the mean scores ($\beta_3 = 0.215 \pm 0.068$, p = 0.034). In table 3 values for the baseline trend and changes after STARD statement publication is provided.

DISCUSSIONS

With the publication of many diagnostic studies in medical journals, it has become quite important to adhere to publishing standards like STARD, CONSORT, and STROBE etc. Publishing standards allow us to establish a benchmark against which every published article can measure up. In this study, we have tried to measure the actual success of a publishing standard (STARD) in improving the reporting quality of diagnostic studies. For this purpose, we used a major medical journal IJMR which has a long illustrious history among medical journals.

Several studies have previously studied the impact of reporting guidelines/statements like CONSORT, STARD or STROBE. All of them have suggested that using the statement might improve the overall reporting of published studies [Moher et al., 2001; Hopewell et al., 2010; Kane et al., 2007]. However, all these studies usually use the uncontrolled version of before-after study design. Previous published evidence have shown that such uncontrolled before-after analysis which tends to compare a pre-and post-time around an intervention may in turn lead us to overestimate the effect of the said intervention [Eccles et al., 2003]. To take into this account, we used an interrupted time-series analyses. It is considered a very powerful statistical method for distinguishing the underlying trend from the actual effects of a given intervention [Hopewell et al., 2010; Kane et al., 2007, Lopez et al., 2017]. Hence, a well-designed time series analysis has the potential to increase the confidence with which the effect estimate can be ascribed to the intervention in question. This however has a drawback as we cannot separate any other effects which might occur at the same time as the intervention [Eccles et al., 2003; Ramsay et al., 2003]. The one major factor which can improve the quality of interrupted-time series analyses is the number of data points collected before and after intervention [Hopewell et al., 2010; Kane et al., 2007]. In the present study, pre-and post-STARD period both have sufficient data points in accordance with the recommendations from Cochrane Effective Practice and Organization of Care group [Moher et al., 2001].

CONCLUSIONS

We conclude that STARD checklist had a statistically significant impact on the reporting quality of diagnostic studies published in India. Our results show that this general improvement would in general lead to better reporting quality of diagnostic accuracy studies if STARD is made an important part of the article submission process in Indian journals. STARD checklist and its extensions, provide a vital tool for researchers not only to use as a guideline for proper reporting but also to conduct diagnostic studies.

We feel, there is a need to continuously educate the medical science professionals regarding formulating research questions properly using correct statistical techniques and reporting required results including testing the validity of assumptions of those techniques.

REFERENCES

1. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981; 94: 557–92.

2. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8: 283–98.

3. Sackett DL, Haynes RB, Guyatt GH, Tugwell T. The selection of diagnostic tests. In: Clinical Epidemiology. A Basic Science for Clinical Medicine. 2nd ed. London: Little, Brown, 1991; 47–57.

4. Lijmer JG, Mol BW, Heisterkamp S et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282: 1061–6

5. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAM A 1995;274:645–51.

6. 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.

7. 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.

8. Turner L, Shamseer L, Altman DG, et al. Consolidated standardsof reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev 2012;11:MR000030.

9. Bossuyt PM. STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiolog y 2008;248:713–14.

10. 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.

11. Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication—before-and-after study. Radiology 2008;248:817–23.

12. Shrout PE, Fleiss JL. (1979) Intraclass correlations: uses in assessing reliability. Psychol Bull. 86:420–8

13. Eccles M, Grimshaw J, Campbell M, Ramsay C. (2003) Research designs for studies evaluating the effectiveness of change and improvement strategies. Qual Saf Health Care 12: 47–52.

14. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. (2003) Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care 19: 613–623.

15. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. (2002) Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther 27: 299–309

16. Group: CEPaOoc. epoc.cochrane website. Available: http://epoc.cochrane.org/sites/epoc.cochrane.org/files/uploads/21%20Interrupted%20time%20series%20analyses%202013%2008%2012_1.pdf. Published: December 12, 2009. Updated: October 15, 2014. Accessed July 29, 2017.

17. Moher D, Jones A, Lepage L. (2001) Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 285: 1992–1995.

18. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. (2010) The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ 340: c723.

19. Kane RL, Wang J, Garrard J. (2007) Reporting in randomized clinical trials improved after adoption of the CONSORT statement. J Clin Epidemiol. 60:241–249

20. Lopez Bernal J., Cummins S., Gasparrini A. 2016. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology:dyw098. DOI: 10.1093/ije/dyw098

Table 1(on next page)

STARD checklist for the reporting of studies of diagnostic accuracy

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3413v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Section and Topic	**Item Number**	**Description**
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading ’sensitivity and specificity’)
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.
METHODS		
Participants	3	Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.
	4	Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	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
	6	Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?

	7	Was the sampling adequate?
	8	Sampling size calculation was done
Test methods	9	Describe the reference standard and its rationale
	10	Describe technical specifications of material and methods involved including howand when measurements were taken, and/or cite references for index tests and reference standard
	11	Describe definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard
	12	Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard
	13	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.
Statistical Methods	14	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95%confidence intervals)
	15	Describe methods for calculating test reproducibility, if done
	16	16. State which software was used for analysis
RESULTS		
Participants	17	Report when study was done, including beginning and

		ending dates of recruitment
	18	Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers)
	19	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).
Test Results	20	Report time interval from the index tests to the reference standard, and any treatment administered between.
	21	Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.
	22	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.
	23	Report any adverse events from performing the index tests or the reference standard.
Estimates	24	Report estimates of diagnostic accuracy and measures of statistical uncertainty

		(e.g. 95%confidence intervals)
	25	Report how indeterminate results, missing responses and outliers of the index tests were handled
	26	Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done
	27	Report estimates of test reproducibility, if done.
DISCUSSION	28	Discuss the clinical applicability of the study findings

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3413v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Table 2(on next page)

Comparison of correctly reported items between two periods (pre and post-STARD)

Correct use of *n* (%): for each item, *n* is the number of articles reporting the item correctly and the percentage = *n*/the number of papers reporting the items × 100%; for each, *n* is the number of articles with the reported item and the percentage = *n*/ the number of papers reporting the items × 100%. For cells with no value in chi-square column, the p-value was obtained via Fisher's Test

Items in checklist	Correct use pre-STARD n(%)	Correct use post-STARD n(%)	X^2	P-value
TITLE/ABSTRACT/ KEYWORDS				
1. Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity')	10(29%)	17(42.5%)	1.572	0.2099
INTRODUCTION				
2. State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	15(42.9%)	18(44%)	0.008	0.927
METHODS: Participants				
3. Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	1(2.9%)	6(14.6%)	-	0.1158
4. Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	13(42%)	11(27.5%)	0.172	0.6784
5. 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	1(3.57%)	9(25.7%)	5.712	0.0169
6. Was the sampling adequate?	0(0%)	3(7.3%)	-	0.2447
7. Sampling size calculation was done	0(0%)	1(2.44%)	-	NA

8. Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	17(48.57%)	17(42%)	0.386	0.5345
METHODS: Test methods				
9. Describe the reference standard and its rationale	13(37%)	18(44%)	0.357	0.55
10. 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	7(20%)	2(4.9%)	-	0.0719
11. Describe definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard	0(0%)	0(0%)	-	NA
12. Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard	0(0%)	0(0%)	-	NA
13. 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.	-	0(0%)	-	NA
METHODS: Statistical Methods				
14. Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95%confidence	1(2.9%)	6(14.7%)	0.029	0.1158
15. Describe methods for calculating test reproducibility, if done	0(0%)	2(50%)	-	0.4286
16. State which software was used for analysis	0(0%)	10(24.4%)	-	0.0014

RESULTS: Participants				
17. Report when study was done, including beginning and ending dates of recruitment	0(0%)	0(0%)	-	NA
18. Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers)	1(2.9%)	7(17%)	-	0.0627
19. 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).	0(0%)	0(0%)	-	NA
RESULTS: Test Results				
20. Report time interval from the index tests to the reference standard, and any treatment administered between.	0(0%)	0(0%)	-	NA
21. Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	0(0%)	1(2.7%)	-	NA
22. 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.	2(6.25%)	14(34.14%)	-	0.0045
23. Report any adverse events from performing the index tests or the	0(0%)	0(0%)	-	NA

reference standard.				

RESULTS: Estimates				
24. Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95%confidence intervals)	1(2.85%)	6(14.6%)	-	NA
25. Report how indeterminate results, missing responses and outliers of the index tests were handled	0(0%)	0(0%)	-	NA
26. Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done	0(0%)	0(0%)	-	NA
27. Report estimates of test reproducibility, if done	0(0%)	0(0%)	-	NA
DISCUSSION				
28. Discuss the clinical applicability of the study findings	6(17.4%)	22(54%)	9.704	0.0018

Table 3(on next page)

Parameter estimates from the Interrupted time-series model predicting the mean yearly score per article

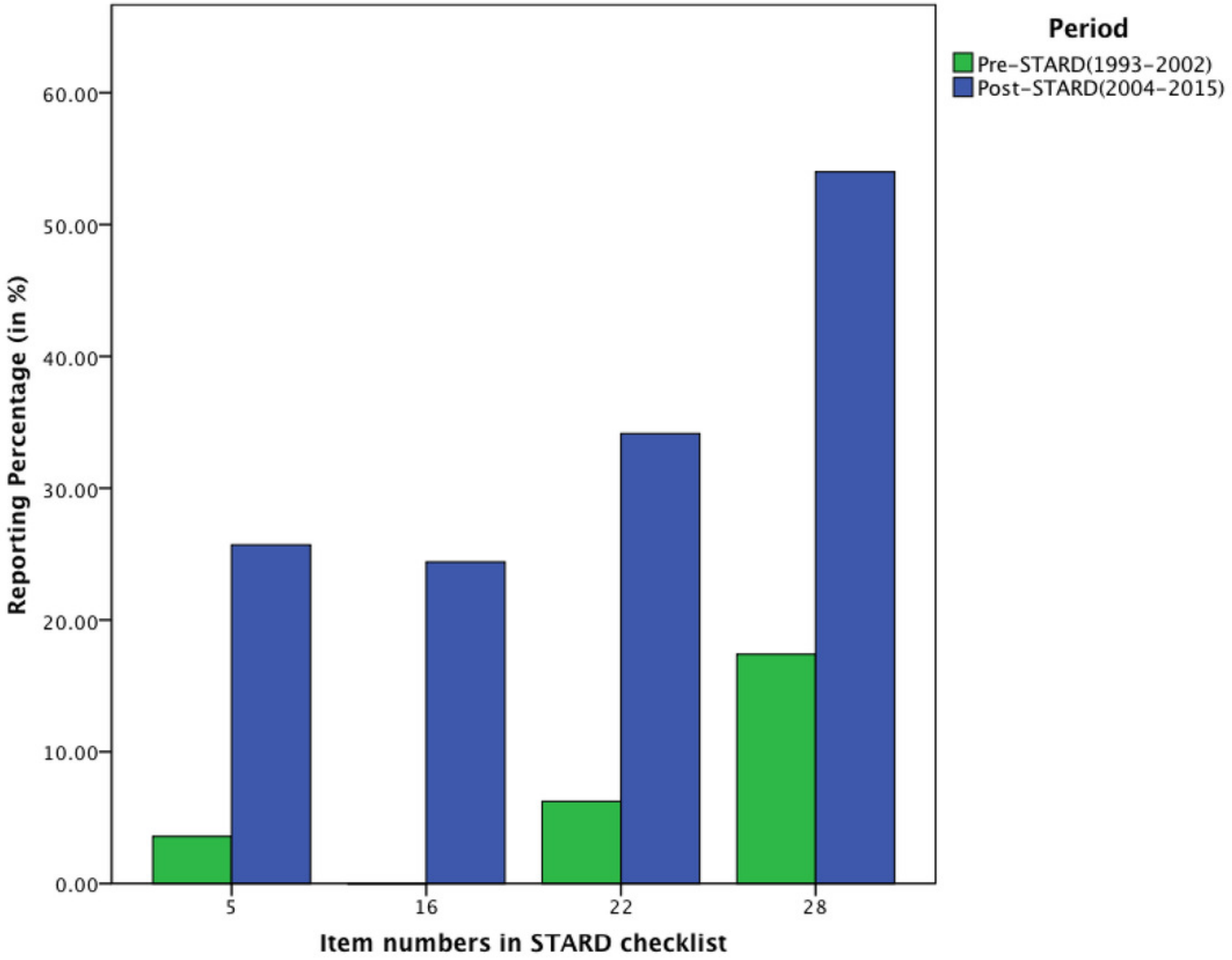
	Estimated Coefficient (Standard deviation)	**P-value**
Interrupted time-series model		
1st segment (pre-STARD, 1993 to 2002)		
Intercept	0.155(0.026)	0.004
Baseline Trend	0.038(0.011)	0.028
2nd segment (post- STARD, 2004 to 2013)		
Trend Change	0.215(0.068)	0.034

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3413v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Figure 1

Comparison of certain scores before and after STARD publication (1993-2013)

This figure gives a gist of the average mean score of all articles for certain selected items for which a significant change was observed in post-STARD period


Period
🟩 Pre-STARD(1993-2002)
🟦 Post-STARD(2004-2015)
Reporting Percentage (in %)
Item numbers in STARD checklist