DOI:
10.1016/j.jebdp.2019.02.002

Document Version
Peer reviewed version

Title

Adequate reporting of dental Diagnostic Accuracy Studies is lacking: an assessment of reporting in relation to the STARD statement

Authors

Michael Durkan[a], Ravi Chauhan[b], Nikolaos Pandis[c], Martyn T. Cobourne[a], Jadbinder Seehra[a*]

Affiliations

[a] Department of Orthodontics, Faculty of Dentistry, Oral and Craniofacial Sciences King's College London, Floor 22, Guy's Hospital, Guy's and St Thomas NHS Foundation Trust, London, SE1 9RT, United Kingdom

[b] Department of Restorative Dentistry, King's College London Dental Institute, Kings College Hospital NHS Foundation Trust, Bessemer Road, London, SE5 9RS, United Kingdom

[c] Department of Orthodontics and Dentofacial Orthopedics, Dental School/Medical Faculty, University of Bern, Freiburgstrasse 7 CH-3010, Bern, Switzerland

***Author for correspondence**

Mr Jadbinder Seehra
Department of Orthodontics
Faculty of Dentistry, Oral and Craniofacial Sciences
King's College London
Floor 22
Guy's Hospital
Guy's and St Thomas NHS Foundation Trust
London
SE1 9RT
United Kingdom
Email:jadbinderpal.seehra@kcl.ac.uk
Telephone number: 02071885665

Title

Adequate reporting of dental Diagnostic Accuracy Studies is lacking: an assessment of reporting in relation to the STARD statement

Abstract

Objective: To assess the quality of reporting of full text articles of dental diagnostic accuracy studies published in eight leading specialty dental journals in relation to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement.

Methods: The full articles of all included studies were assessed for their adherence to the 30 item STARD checklist by two researchers independently. A score of 0-2 was attributed to each item. Inter-rater agreement was assessed. Univariate and Multivariate linear regression was carried out to evaluate differences in reporting qualities between journals and whether certain variables influenced reporting qualities.

Results: 145 articles were identified. Full article STARD checklist items relating to methodology and results were poorly reported. The overall mean quality score for full articles was 28.75. Articles published in the Journal of Cranio-maxillo-facial surgery obtained the highest quality score. In the multivariate analysis, articles published in Journal of Cranio-maxillo-facial surgery had significantly higher reporting quality scores compared to those published in European Journal of Orthodontics (β =-6.97, 95% CI: -11.62, -2.30, $p<0.05$), Journal of Prosthetic Dentistry (β =-8.01, 95% CI: -14.60, -1.41, $p<0.05$) and Oral Diseases (β =-6.72, 95% CI: -11.57, -1.86, $p<0.05$). Reporting quality improved per year ($p<0.028$).

Conclusion: Adherence to STARD for full articles is suboptimal in dental journals.

Keywords: Evidence-based dentistry, Dentistry, Diagnostic, STARD statement,

Running title: Diagnostic Accuracy Studies and the STARD statement

Word count: 3000

Introduction

The Standards for Reporting of Diagnostic Accuracy Studies (STARD) aims to improve the inadequate reporting of diagnostic accuracy studies. It assists researchers in their preparation of study reports and assists readers in assessing the quality/completeness of reporting of diagnostic accuracy studies. Diagnostic accuracy studies assess the ability of a diagnostic test to identify the presence or absence of a target condition by comparing their relative diagnostic strength with a reference standard. Errors in the methodology of diagnostic studies can result in systematic bias of the estimates of accuracy. [1-3] The purpose of a diagnostic test is to assess whether there is the presence or absence of disease. 'Test' is the term used to describe any procedure used 'to gather information on the health status of an individual'. [4] With the plethora of available diagnostic tests, it is crucial that the clinicians are able to identify and select the most accurate test for use in their cohort of patients. This will require the clinician to possess the appropriate skills to critically appraise the literature, but also requires clear standardized reporting of diagnostic studies. Exaggerated and biased results that are reported in diagnostic accuracy studies may result in the uptake of such tests into clinical practice, leading to incorrect diagnosis and management together with an increase in healthcare costs associated with inappropriate tests.[5]

In 1999, the diagnostic and screening test methods working group identified and discussed the low methodological quality and poor reporting of diagnostic accuracy studies. It was hoped that the production of a checklist would address the issues of poor reporting similarly to CONSORT for randomised controlled trials. [6-8] The STARD checklist was published in 8 medical journals and by 2008 it was estimated that over 200 biomedical journals encouraged the use of STARD in their instructions to authors.

Korevaar et al., [9] evaluated the adherence of diagnostic accuracy studies published in 2012 to the STARD statement and whether there was a difference in reporting qualities compared with studies published in 2000 and 2004. It was concluded that there was an improvement in the reporting 10 years after the introduction of STARD (on average, an improvement of 3.4 items and 1.7 items compared with studies published in 2000 and 2004 respectively). However, the completeness of reporting remains suboptimal for many articles. Poorly reported items included: inclusion criteria, sampling methods for recruiting patients, information about blinding and confidence intervals for accuracy estimates. [9]

In 2013, the STARD 25-item list for full text articles was updated to include recent evidence on sources of bias and addressing applicability concerns and factors allowing the over-optimistic interpretation of diagnostic accuracy tests. [10] Further items were introduced resulting in an increase in the number of checklist items from 25 to 30 (Appendix I). A recommendation regarding the inclusion of an accompanying flowchart that describes the flow of patients through the study was also included. [10] There have been several studies assessing the quality of reporting of full article diagnostic accuracy studies in medicine using the STARD statement. [11-24]

These have shown that there are certain items that are commonly under-reported (Table 1). When providing healthcare treatment and management it is crucial that dental clinicians can identify and select the most accurate test for use in their cohort of patients. This will require the clinician to possess the appropriate skills to critically appraise the literature, but also requires clear standardized reporting of dental diagnostic studies. To our knowledge there are no studies within the dental literature assessing the quality of reporting of diagnostic accuracy studies. The aim of this study was to assess the quality of reporting of full text articles of dental diagnostic accuracy studies published in eight leading specialty dental journals in relation to the STARD statement for full article texts.

Materials and Methods

A decision was made to review the quality of reporting of diagnostic accuracy studies in dental journals from various disciplines. The highest rating impact factor journals in the field of dentistry were sourced from the Thomson Reuters Web of Science database in 2014 (http://dentistry.sbmu.ac.ir/uploads/2014.DentalJournalsImpactFactor.pdf). Eight leading dental specialty journals (Journal of Clinical Periodontology (JCP), Clinical Oral Implants Research (COIR), Journal of Endodontics (JOE), European Journal of Orthodontics (EJO), Journal of Cranio-maxillo-facial surgery (JCMFS), Paediatric Dentistry (PD), Journal of Prosthetic Dentistry (JPD) and Oral Diseases (OD)) were identified based on their impact factor. A literature search of these journals was carried out to identify all diagnostic accuracy studies published in these journals between 2003-2015. Electronic searching was carried out using a search filter in PubMed with high sensitivity for diagnostic accuracy studies ("sensitivity AND specificity" [MH] OR specifit* [TW] or "false negative" [TW] or accuracy [TW]). [19,25]

Inclusion and exclusion criteria

Studies were selected for inclusion if the abstract reported estimates of the accuracy of the test on humans and was based on a comparison of index test results against a clinical reference standard. Review articles, meta-analyses, systematic reviews, longitudinal studies, letters, conference abstracts and laboratory studies involving extracted human teeth were excluded. However, cadaver studies, including teeth that were in extracted mandibles were included.

Two researchers (MD and RC) screened all potentially relevant abstracts independently. Any disagreements were resolved by discussion with a third author (JS) to reach a consensus. Both researchers (MD and RC) were calibrated by assessing the reporting of 5 full text articles. During the assessment of full text articles, both authors also referred directly to the STARD checklist for full text articles and their associated explanation. [10] The levels of

agreement between the authors (MD and RC) for full text articles were assessed using Kappa statistics. The STARD 2015 checklist was used to evaluate the quality of reporting of each full text article. [10] The STARD checklist consists of 30 items. The score for each item ranged from 0 to 2, with 0 representing 'no reporting', a score of 1 corresponding to 'partial reporting' and a score of 2 corresponding to 'full reporting'. Similar ratings have been previously employed to calculate overall quality scores in epidemiological studies assessing the quality of the reporting of the abstracts of randomised controlled trials. [27] Both researchers using a pre-specified data collection sheet extracted all data independently. The scores for the items from both researchers were assessed and combined. A percentage score was calculated for the full text articles with a score of 68 for full text articles representing perfect reporting. As per previous studies [27], to determine factors which may predict adherence and reporting to the STARD checklist the following variables were collected: journal of publication, number of authors, type of study, continent of publication, significance of results and identification of the study as either a single or multicentre study were also extracted.

Inter-rater reliability was calculated using Kappa coefficient. Descriptive statistics (not reported, partially reported, fully reported) for individual reporting items were calculated and converted to a percentage scale. Mean quality scores for each full text article was calculated. Implementing linear regression modelling with a univariate analysis identified characteristics associated with the mean score. The adjusted effect on reporting quality score was identified by multivariate modeling. The multivariable model was employed to test significant predictors identified by the univariable model. A two-tailed p-value of 0.05 was considered statistically significant. Statistical analyses were performed with STATA® version 14.2 software (Stata Corporation, College Station, Texas, USA).

Results

Inter-rater reliability was assessed using the Kappa coefficient. The overall inter-rater reliability for both assessors (MD and RC) was at a good level with a score of 0.674. A total of 145 full text articles were identified (Appendix.II). Figure 1 shows the entire search process. Table 2 reports the characteristics of the included studies. Within the included sample, 20% (n=29) of diagnostic accuracy studies were obtained from the Journal of Endodontics whilst 4.2% (n=6) were provided by the Journal of Prosthetic Dentistry. The majority of articles were published in Europe (48.3%), single gate type (82.8%) and single centre (98.6%). The level of significance was reported in 56.5% (n=82) of the articles . The mean number of authors was 5.09 (range 2-14). The number of articles published per year is shown in Figure 2.

Fully reported items included: Scientific and clinical background, including the intended use and clinical role of the index test (99.7%), Study objectives and hypotheses (88.3%), Index test, in sufficient detail to allow replication (80%), and Implications for practice, including the intended use and clinical role of the index test (95.8%). Very poorly reported items included how indeterminate index test or reference standard results were handled (2.1%), how missing data on the index test and reference standard were handled (4.8%), intended sample size and how it was determined (6.9%), flow of participants, using a diagram (4.1%) and any adverse events from performing the index test or the reference standard (2.7%) (Table 3).

The overall mean quality score for full text articles was 28.75. The highest quality score was obtained by articles published in the Journal of Cranio-maxillo-facial surgery with a mean score of 33.8 (SD 7.5). In contrast, the lowest quality score for articles was obtained by European Journal of Orthodontics with a mean score of 24.3 (SD 6.9) (Table 4).

Comparisons were made between the baseline (reference category) and the following predictors: individual journals, number of authors, continent of publication, study type, single-

or multi-centre studies, publication year and the presence of significant and non-significant findings (Table 5). The univariate model revealed that the Journal of Cranio-maxillo-facial surgery had higher quality scores than all the other journals, with a statistically significant higher quality score compared to Clinical Oral Implants Research (β =-5.38, 95% CI: -10.15, -0.61, p<0.05), European Journal of Orthodontics (β =-9.49, 95% CI: -13.74, -5.23, p<0.05), Journal of Prosthetic Dentistry (β =-9.28, 95% CI: -15.71, -2.85, p<0.05) and Oral Diseases (β =-7.01, 95% CI: -11.63, -2.40, p<0.05).

It was noted that there was a trend for the quality of reporting, in terms of adherence to STARD, to improve as the years progressed (p<0.000) and the numbers of authors increased (p<0.003). The multivariate analysis demonstrated a similar trend with evidence that the quality of reporting in full articles of diagnostic studies in Journal of Cranio-maxillo-facial surgery were better than all other journals. This was statistically significant in comparison with European Journal of Orthodontics (β =-6.97, 95% CI: -11.62, -2.30, p<0.05), Journal of Prosthetic Dentistry (β =-8.01, 95% CI: -14.60, -1.41, p<0.05) and Oral Diseases (β =-6.72, 95% CI: -11.57, -1.86, p<0.05). It also revealed that there was a statistically significant improvement in mean quality scores per year (p<0.028). There was also a trend for the quality of reporting to be better in studies published in America and Asia. Similarly, multiple gate studies were associated with higher reporting qualities.

Discussion

The quality of reporting of full articles of diagnostic accuracy studies in relation to the STARD checklist is suboptimal. The overall mean quality score for full articles was 28.75. Variation in the reporting quality between dental speciality journals was also recorded. Articles published in the Journal of Cranio-maxillo-facial surgery had significantly higher reporting quality scores compared to other specialty journals. Similarities can be drawn between our study results and the findings of previous studies (Table 1) investigating the reporting and adherence of medical diagnostic studies to the STARD statement for full-text articles. Korevaar et al., [9] found items such as inclusion criteria, sampling methods for recruiting patients, information about blinding and confidence intervals for accuracy estimates to be sub-optimally reported. Within our sample similar trends are evident. Poorly reported items are include key items relating to methodology, results and conclusions. Under-reported methodological items include those describing blinding, time interval between the index and reference test and the reporting of the distribution of the severity of the disease or alternative diagnoses in participants without the target condition. Reporting of the handling of indeterminate, missing results, baseline demographic and clinical characteristics of participants, whether clinical information and index test results were available to the assessors of the reference standard, determination of sample size and discussion of potential adverse events were also suboptimal. It is concerning that these items were fully reported in equal to or less than 15% of the dental diagnostic accuracy studies included in this study given their potential to bias the results. Reviewer bias can arise with knowledge of index test results prior to reading the reference standard test, thus demonstrating the importance of blinding in a study. The lack of knowledge on the extent of indeterminate and missing results can have a significant impact on the estimates of accuracy and these should always be reported in the full text. Insufficient reporting of baseline demographics prohibits the reader from making an assessment of sampling bias. No mention of time intervals between tests can lead to treatment paradox and disease progression bias. Adverse effects

should be discussed so as to give the clinician an insight into the suitability of applying a test to a patient because it may mean that the test should be avoided irrespective of its accuracy. This study has also shown that there was a statistically significant increase in the mean quality score per year between 2003-2015 ($P<0.028$). This is similar to previous findings where both a mean increase of 0.3 per item [14] and a significant increase in the mean quality score per year [26] was reported suggesting an increased awareness and adoption of STARD checklist.

Strengths and weakness

Every effort was made to reduce bias in the study in an attempt to improve the generalizability of the results. Selection bias was reduced by adhering to the definition of "diagnostic studies" that satisfied the inclusion/exclusion criteria. In addition, screening of relevant studies was undertaken independently by two authors with all disagreements resolved by discussion with a third author to reach a consensus. Inter-rater variability was reduced by calibration, direct referral to the STARD checklist and their associated explanation when assessing the articles and a good level of inter-rater agreement assessed using the Kappa coefficient (0.674). The decision was made to include studies from various disciplines within the field of dentistry because it would allow both a more balanced examination of the overall quality of reporting of diagnostic accuracy studies in dentistry. The literature search was carried out between the years of 2003 and 2015 so as to only review the quality of reporting after the introduction of the STARD checklist in 2003 and the updated version in 2015.

We felt that it was appropriate not to include diagnostic studies that were based on extracted teeth as these maybe unrepresentative of the true performance of a test in vivo. We felt this would help to avoid spectrum bias as the sample under investigation would not be representative of our sample of patients that we see in the clinical environment. Our decision to include studies that exhibited teeth remaining in extracted mandibles was based

on the opinion that this more closely reflected the true clinical situation. This was because many of the diagnostic tests examining new radiological techniques/equipment used extracted mandibles. We felt that the preservation of the hard bony envelope surrounding the teeth offered a more realistic view of the clinical situation.

We examined all of the items on the STARD checklist. We did not make amendments for non-applicable items. For instance, there are items (item 21) relating to disease prevalence. The first part pertains to 'Distribution of severity of disease in those with the target condition' and the second part refers to 'distribution of alternative diagnoses in those without the target condition'. With many of the studies that we evaluated, this item wasn't applicable. For example in radiological diagnostic accuracy tests, some studies examined the accuracy of the test in locating cephalometric points. In these circumstances, there is no disease and as such the study will not score positively for the item. Similarly, these items do not apply to the many studies in the Journal of Endodontics, where the accuracy of various apex locators are compared. An example of another item, which was not always applicable, is item 25 regarding 'adverse effects'. While this item is clearly applicable for studies having a potential adverse effect on humans such as radiographic studies, it was not applicable for studies assessing the root length of teeth in extracted mandibles.

Including non-applicable items has the effect of giving a study a lower STARD score, which may not fairly reflect the quality of reporting in that particular study. Other studies examining adherence to STARD in full articles, such as Wilczynski [14] used a modified version of the STARD checklist (13 out of 25 items) in order to avoid this potential error. However, in a cross discipline study such as this, it is difficult to create a modified STARD checklist that can be applied to the entire sample.

Implications for further research

An assessment of the reporting between various specialities in dentistry was beyond the remit of this study but may be considered as an aim in future studies. Further research is required with a modified STARD statement that considers non-applicable items and is specialty specific in order to preserve consistency. This could be considered in future studies assessing dental diagnostic studies with the STARD checklist allowing a fairer representation of the quality of reporting.

Conclusion

This study has identified that the reporting of full text articles of dental diagnostic studies in relation to the STARD checklist requires improvement. Despite an improvement in the mean quality scores per year, variation in the reporting quality between specialty journals is evident.

References

1. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999 ;282 :1061-6.
2. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006 ;174 :469-76.
3. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol. 2005 Jun 8;5:20. Review.
4. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 10;3:25.
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003 ;138: 1-12.
6. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996 ;276 :637-9.
7. Moher D, Jones A, Lepage L; CONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of

randomized trials: a comparative before-and-after evaluation. JAMA. 2001;285 :1992-5.

8. Moher D, Schulz KF, Altman D; CONSORT Group (Consolidated Standards of Reporting Trials).The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285 :1987-91.
9. Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting diagnostic accuracy studies:some improvements after 10 years of STARD. Radiology. 2015;274 :781-9.
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015; 277:826-32.
11. Coppus SF, van der Veen F, Bossuyt PM, Mol BW. 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.
12. Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS One. 2009 ;4: e7753.
13. Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. Phys Ther. 2001;81:1546-64.
14. Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. Radiology. 2008 ;248 :817-23.
15. Johnston KC, Holloway RG. There is nothing staid about STARD: progress in the reporting of diagnostic accuracy studies. Neurology. 2006;67:740-1. No abstract available. Erratum in: Neurology. 2008 Jul 8;71(2):152.
16. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al. Quality of reporting of diagnostic accuracy studies. Radiology. 2005 ;235 :347-53. Epub 2005 Mar 15.

17. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology. 2006 ;67 :792-7.
18. Walther S, Schueler S, Tackmann R, Schuetz GM, Schlattmann P, Dewey M. Compliance with STARD checklist among studies of coronary CT angiography: systematic review. Radiology. 2014 ;271 :74-86.
19. Korevaar DA, Cohen JF, Hooft L, Bossuyt PM. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies. J Clin Epidemiol. 2015 ;68 :708-15. doi: 10.1016/j.jclinepi.2015.01.014. Epub 2015 Jan 28. Review.
20. Maclean EN, Stone IS, Ceelen F, Garcia-Albeniz X, Sommer WH, Petersen SE. Reporting standards in cardiac MRI, CT, and SPECT diagnostic accuracy studies: analysis of the impact of STARD criteria. Eur Heart J Cardiovasc Imaging. 2014;15 :691-700.
21. Rama KR, Poovali S, Apsingi S.
Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal.
Clin Orthop Relat Res. 2006; 447:237-46
22. 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.
23. Zafar A, Khan GI, Siddiqui MA. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. Clin Exp Ophthalmol. 2008 ;36 :537-42.
24. Zintzaras E, Papathanasiou AA, Ziogas DC, Voulgarelis M. 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.

25. Devillé WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol. 2000 ;53 :65-9.
26. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med. 2014 ;19 :47-54.
27. Seehra J, Wright NS, Polychronopoulou A, Cobourne MT, Pandis N. Reporting quality of abstracts of randomized controlled trials published in dental specialty journals. J Evid Based Dent Pract. 2013 ;13:1-8.

Legends to tables and figures

Table 1. Underreported STARD checklist items in medical journals.

Table 2. Characteristics of diagnostic studies included (n=145)

Table 3. Percentage distribution of scoring for each item in the 30-item STARD checklist for full text articles (n= 145)

Table 4. Quality assessment score of the 145 diagnostic studies (full articles) per journal type

Table 5. Univariate and multivariate linear regression derived coefficients (b) and 95% confidence intervals (CI) for quality assessment (score) as dependent variable for the 145 full text articles.

Figure 1. Flowchart demonstrating study selection

Figure 2. Frequency of publications (per year) included

Appendix I STARD Checklist for full text articles