Reporting standards in cardiac MRI, CT, and SPECT diagnostic accuracy studies: analysis of the impact of STARD criteria.

Maclean, EN; Stone, IS; Ceelen, F; Garcia-Albeniz, X; Sommer, WH; Petersen, SE

European Heart Journal - Cardiovascular Imaging

Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria

--Manuscript Draft--

Manuscript Number:	EHJCI-D-13-00703R2
Full Title:	Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria
Article Type:	Original Paper
Keywords:	Diagnostic accuracy, STARD, reporting quality
Corresponding Author:	Steffen E Petersen, M.D. NIHR Cardiovascular Biomedical Research Unit at Barts London, UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	NIHR Cardiovascular Biomedical Research Unit at Barts
Corresponding Author's Secondary Institution:	
First Author:	Edd N Maclean, MBBS BSc AKC
First Author Secondary Information:	
Order of Authors:	Edd N Maclean, MBBS BSc AKC
	Ian S Stone, MD MBCHB
	Felix Ceelen
	Xabier Garcia-Albeniz, MD PHD MPH
	Wieland H Sommer, MD
	Steffen E Petersen, M.D.
Order of Authors Secondary Information:	
Abstract:	Aims: Diagnostic accuracy studies determine the clinical value of non-invasive cardiac imaging tests. The 'Standards for the reporting of diagnostic accuracy studies' (STARD) were published in 2003 to improve the quality of study reporting. We aimed to assess the reporting quality of cardiac computed tomography (CCT), single positron emission computed tomography (SPECT) and cardiac magnetic resonance (CMR) diagnostic accuracy studies; to evaluate the impact of STARD; and to investigate the relationships between reporting quality, journal impact factor and study citation index. Methods and Results: We randomly generated 6 groups of 50 diagnostic accuracy studies: 'CMR 1995-2002', 'CMR 2004-2011', 'CCT 1995-2002', 'CCT 2004-2011', 'SPECT 1995-2002', and 'SPECT 2004-2011'. The 300 studies were double-read by 2 blinded reviewers and reporting quality determined by % adherence to the 25 STARD criteria. Reporting quality increased from 65.3% before STARD to 74.1% after ($p=0.003$) in CMR studies and from 61.6% to 79.0% ($p<0.001$) in CCT studies. SPECT studies showed no significant change: 71.9% before and 71.5% after STARD ($p=0.92$). Journals advising authors to refer to STARD had significantly higher impact factors than those that did not ($p=0.03$), and journals with above-median impact factors published studies of significantly higher reporting quality ($p<0.001$). Since STARD, citation index has not significantly increased ($p=0.14$) but, after adjustment for impact factor, reporting quality continues to increase by approximately 1.5% each year. Conclusion: Reporting standards for diagnostic accuracy studies of non-invasive cardiac imaging are at most satisfactory and have improved since the introduction of STARD. Adherence to STARD should be mandatory for authors of diagnostic accuracy studies.

Barts and The London
School of Medicine and Dentistry

WILLIAM HARVEY RESEARCH INSTITUTE

Steffen E Petersen
Professor of Cardiovascular Medicine
Centre Lead

Centre for Advanced Cardiovascular Imaging
William Harvey Research Institute
Barts and The London
NIHR Biomedical Research Unit
The London Chest Hospital
Bonner Road
London E2 9JX, UK

s.e.petersen@qmul.ac.uk

Gerald Maurer, MD
Editor-in-Chief
European Heart Journal - Cardiovascular Imaging

December 12, 2013

RE: Re-resubmission Ms. No. EHJCI-D-13-00703: Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria

Dear Professor Maurer,

On behalf of all co-authors I thank you for the opportunity to resubmit our re-revised manuscript entitled "Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria". The reviewers' feedback was very thoughtful and we hope to have addressed all of the comments appropriately in the enclosed rebuttal letter and in the revised manuscript (one with and one without tracked changes). We feel that this has substantially improved the quality of the manuscript and hope that the manuscript is now acceptable for publication in this journal.

This manuscript is not under consideration for publication elsewhere. None of the paper"s contents have been previously published. All authors have read and approved the manuscript. There are no potential conflicts of interest for any of the authors.

I look forward to hearing from you in due course.

Yours sincerely,

Steffen E Petersen, MD DPHIL MPH FRCP FESC FACC
Professor of Cardiovascular Medicine

EHJCI-D-13-00703: Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria

RESPONSE

We very much appreciate the reviewer's ongoing input and his kind comments regarding the manuscript. We hope we have sufficiently addressed his suggestions below:

1) Please consider whether you want to revise the result section as following: "? when compared to those that did not (81.2 ±9.3% vs. 74.3 ± 12.7%, respectively; p=0.06). However, after adjusting for potential confounder including impact factor and year of publication, this difference attenuated (p=0.72)."

We thank the reviewer for suggesting clearer articulation and have adjusted the body of text accordingly: "*However, after adjusting for the potential confounder impact factor and year of publication, this difference attenuated (p=0.72)."*

2) I guess there are also some typos in the results section. Without having the actual stats printouts, is this version correct? "?before the introduction of STARD (estimate: -0.018, p=0.97). However, during the 8 years after the introduction of STARD, the model suggests an increase in adherence to STARD criteria by 1.46% (sum of estimates: -0.018+1.479=1.461, p=0.04) for each calendar year (Figure 2)."

The confusion and discrepancy is likely due to rounding issues. To make it clearer we rounded to the same number of decimals and have adjusted our manuscript accordingly. "*After controlling for the potential confounder impact factor, reporting quality remained unchanged in the 8 years before the introduction of STARD (estimate -0.018, p=0.97). However, during the 8 years after the introduction of STARD, the model suggests an increase in adherence to STARD criteria by 1.461% (sum of estimates -0.018 +1.479, p=0.04) for each calendar year (Figure 2)."*

3) Please check again your way reporting p-values, particular in table 2. I strongly recommend to round all p-values >=0.01 to two decimal places (example 0.025 -> 0.03; 0.013 -> 0.01) and round p-values <0.01 to the first digit which is unequal to zero (example 0.0076 -> 0.008). Why is the p-value for criteria 3 comparing SPECT publications in the revised version now 0.029 and in the original version 0.048?

We have adjusted the p values as recommended throughout the manuscript. We thank the reviewer for identifying the discrepancy regarding criteria 3 in table 2 – this figure appears to be the result of a typographical error that occurred during the recalculation of p values to discern decimal places. The original figure of p=0.048 is correct and has been reinstated and rounded to p=0.05 as recommended. All other p values that were recalculated have been rechecked and remain as written.

4) Please correct the scaling of the x-axis of Figure 2 a and b that they are identical

We have redone both Figures 2a and 2b and ensured they have same x-axis, including scaling and font sizes and symbol shapes and reference line types.

ABSTRACT

Aims: Diagnostic accuracy studies determine the clinical value of non-invasive cardiac imaging tests. The 'Standards for the reporting of diagnostic accuracy studies' (STARD) were published in 2003 to improve the quality of study reporting. We aimed to assess the reporting quality of cardiac computed tomography (CCT), single positron emission computed tomography (SPECT) and cardiac magnetic resonance (CMR) diagnostic accuracy studies; to evaluate the impact of STARD; and to investigate the relationships between reporting quality, journal impact factor and study citation index.

Methods and Results: We randomly generated 6 groups of 50 diagnostic accuracy studies: 'CMR 1995-2002', 'CMR 2004-2011', 'CCT 1995-2002', 'CCT 2004-2011', 'SPECT 1995-2002', and 'SPECT 2004-2011'. The 300 studies were double-read by 2 blinded reviewers and reporting quality determined by % adherence to the 25 STARD criteria. Reporting quality increased from 65.3% before STARD to 74.1% after ($p=0.003$) in CMR studies and from 61.6% to 79.0% ($p<0.001$) in CCT studies. SPECT studies showed no significant change: 71.9% before and 71.5% after STARD ($p=0.92$). Journals advising authors to refer to STARD had significantly higher impact factors than those that did not ($p=0.03$), and journals with above-median impact factors published studies of significantly higher reporting quality ($p<0.001$). Since STARD, citation index has not significantly increased ($p=0.14$) but, after adjustment for impact factor, reporting quality continues to increase by approximately 1.5% each year.

Conclusion: Reporting standards for diagnostic accuracy studies of non-invasive cardiac imaging are at most satisfactory and have improved since the introduction of STARD. Adherence to STARD should be mandatory for authors of diagnostic accuracy studies.

KEYWORDS

Diagnostic accuracy, STARD, reporting quality

Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria

Edd N Maclean[1], Ian S Stone[1], Felix Ceelen[2], Xabier Garcia-Albeniz[3], Wieland H Sommer[2], Steffen E Petersen*[1]

Affiliations: [1]Advanced Cardiovascular Imaging, NIHR Cardiovascular Biomedical Research Unit at Barts, William Harvey Research Institute, Queen Mary University of London, UK; [2] Department of Clinical Radiology, University of Munich, Grosshadern Campus, Munich, Germany; [3] Department of Epidemiology, Harvard School of Public Health, USA.

***Corresponding author: Steffen E. Petersen, MD DPhil FRCP FESC FACC**

Professor of Cardiovascular Medicine,

Honorary Consultant Cardiologist,

Centre Lead for Advanced Cardiovascular Imaging,

William Harvey Research Institute,

NIHR Cardiovascular Biomedical Research Unit at Barts,

The London Chest Hospital,

Bonner Road,

London,

E2 9JX,

UK

Email: s.e.petersen@qmul.ac.uk

Words: ~~2878~~2982

ABSTRACT

Aims: Diagnostic accuracy studies determine the clinical value of non-invasive cardiac imaging tests. The 'Standards for the reporting of diagnostic accuracy studies' (STARD) were published in 2003 to improve the quality of study reporting. We aimed to assess the reporting quality of cardiac computed tomography (CCT), single positron emission computed tomography (SPECT) and cardiac magnetic resonance (CMR) diagnostic accuracy studies; to evaluate the impact of STARD; and to investigate the relationships between reporting quality, journal impact factor and study citation index.

Methods and Results: We randomly generated 6 groups of 50 diagnostic accuracy studies: 'CMR 1995-2002', 'CMR 2004-2011', 'CCT 1995-2002', 'CCT 2004-2011', 'SPECT 1995-2002', and 'SPECT 2004-2011'. The 300 studies were double-read by 2 blinded reviewers and reporting quality determined by % adherence to the 25 STARD criteria. Reporting quality increased from 65.3% before STARD to 74.1% after (p=0.003) in CMR studies and from 61.6% to 79.0% (p<0.001) in CCT studies. SPECT studies showed no significant change: 71.9% before and 71.5% after STARD (p=0.92~~2~~). Journals advising authors to refer to STARD had significantly higher impact factors than those that did not (p=0.~~025~~03), and journals with above-median impact factors published studies of significantly higher reporting quality (p<0.001). Since STARD, citation index has not significantly increased (p=0.~~139~~14) but, after adjustment for impact factor, reporting quality continues to increase by approximately 1.5% each year.

Conclusion: Reporting standards for diagnostic accuracy studies of non-invasive cardiac imaging are at most satisfactory and have improved since the introduction of STARD. Adherence to STARD should be mandatory for authors of diagnostic accuracy studies.

KEYWORDS

Diagnostic accuracy, STARD, reporting quality

ABBREVIATIONS LIST

CMR, cardiac magnetic resonance;
CCT, cardiac computed tomography;
SPECT, single positron emission computed tomography;
STARD, Standards for the reporting of diagnostic accuracy studies;

INTRODUCTION

Advanced non-invasive cardiovascular imaging modalities, such as Cardiovascular Magnetic Resonance (CMR) imaging, Cardiovascular Computed Tomography (CCT) and Single Positron Emission Computed Tomography (SPECT) are increasingly requested clinically. Between 2000 and 2006, Medicare expenditure on medical imaging increased annually by 17% and, since 1996, mean per capita radiation dose has doubled, highlighting the need to avoid unnecessary use of these expensive technologies [1, 2].

Diagnostic accuracy is an important consideration in determining the cost-effectiveness of an imaging test, but often varies amongst different publications. This may reflect the dependence of the results on factors such as study design, patient population and technical considerations as well as random variability. Insufficient reporting may not allow assessment of the internal and external validity of the study findings. Furthermore, over-optimistic diagnostic accuracy results can lead to the premature dissemination of imaging tests and consequently to doctors making incorrect management decisions, contributing to the significant rise in health care costs.

In 2003, the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) group published a set of 25 criteria with the objective of improving the reporting quality of diagnostic accuracy studies [3-5]. These criteria allow the reader to identify the potential for bias in the study (internal validity) and to evaluate whether the results of the studies can be generalized to a wider population (external validity). To date, over 200 journals advise authors to refer to STARD when submitting manuscripts (http://www.stard-statement.org/).

Given the importance of high quality diagnostic test reporting in cardiac patients and the lack of data on adherence to the STARD criteria in this field, the aim of this study was to assess the impact of STARD by comparing the reporting quality of CCT, SPECT and CMR studies published in the eight years before STARD (1995-2002) with those published in the eight years after (2004-2011).

METHODS

Literature Search

A literature search of the Ovid Medline and EMBASE databases was performed on September 1st 2011. We searched for CCT, SPECT and CMR studies of diagnostic accuracy published before (1995-2002 inclusive) and after (2004-2011 inclusive) the introduction of the STARD statement in 2003. The MeSH terms, corresponding number of identified studies and study groupings are shown in Figure 1.

Selection Criteria

We included studies that examined the performance of CCT, SPECT or CMR investigations in relation to a reference standard. Animal models, reviews, meta-analyses, and studies comparing more than one of the CCT, SPECT or CMR modalities against a reference standard were excluded. The identified studies were assigned to one of six groups: CCT, SPECT or CMR diagnostic accuracy studies published before and after the introduction of STARD in 2003, respectively. In each group, eligible studies were numbered and 50 papers selected for analysis using a random number generator (Microsoft Excel 2010, Microsoft). Data extraction was performed from each of the selected abstracts. Studies that did not meet the inclusion criteria during data extraction were excluded and replaced by studies using the random number generator until 50 studies were identified for each of the six groups.

Scoring

The 300 included studies were blinded to date, authorship, references and journal of publication, and read against the criteria of the STARD checklist. Reviewer 1 (E.M.) read all 300 studies; Reviewer 2 (I.S.) read a random selection of 100 studies, and Reviewer 3 (F.C.) read the remaining 200 studies; both reviewers were blinded to the findings of Reviewer 1. Reviewer 4 (S.P.) resolved any disputed decisions. All four reviewers were provided with a document explaining the STARD statement and its rationale, and were instructed to refer to the STARD statement website (www.stard-statement.org) if further clarification on the criteria was required. A total of 7500 STARD items were evaluated within the 300 manuscripts. For each STARD criterion, reviewers assigned 'Yes' if the manuscript addressed

the item appropriately and 'No' if it did not. If a criterion was considered not applicable to the study, such as in retrospective studies where participant dropout does not occur, the abbreviation 'NA' was used.

Impact Factor and Citation Index

Impact factor in the year of study publication was sourced for each journal from the Thomson Reuters Web of Science database. Study Citation Index was calculated by counting the total number of citations in the two years following study publication according to Web of Science data.

Statistics

Reporting quality was assessed in all studies by calculating the percentage adherence to the STARD criteria by dividing the number of agreements with STARD criteria per study by the number of possible agreements (25 criteria minus number of criteria considered not applicable for specific study). We used the following formula:

$$\%\ Adherence = \frac{Number\ of\ 'Yes'}{25 - Number\ of\ 'NA'} \times 100$$

Data were examined for normality (median and mean comparison, skewness, kurtosis, the Shapiro Wilks test and normal probability plots). Normally distributed data are presented as mean ± standard deviation and non-normally distributed data as median (interquartile range). We performed independent t-tests or ANOVA for normally distributed data and the Wilcoxon rank-sum test or Kruskal-Wallis test for independent samples that were not normally distributed. Chi-square tests were used to assess for differences in adherence to all individual STARD criteria.

A linear regression model was built with adherence (%) to STARD criteria as the outcome and the timing of publication with regards to the advent of the STARD criteria in 2003 (before or after STARD) as a binary exposure and potential confounders (impact factor, citation index). A spline with a knot in the year of STARD publication was introduced to allow for a change in the slope. In all cases the significance level was set at $p \leq 0.05$ (two-sided). No adjustment for multiple testing was performed for pre-specified sub-analyses. All statistical analyses were performed using SAS software (Version 9.3; SAS Institute Inc., Cary, NC, US).

RESULTS

Study Selection

As shown in Figure 1, 37 of the initial 300 studies met exclusion criteria and were replaced; 19 were animal studies, 8 were meta-analyses or reviews, 6 studies compared more than one imaging modality to a reference standard, and 4 studies did not examine diagnostic accuracy. Of the included studies, 167 evaluated the diagnostic accuracy of CCT, SPECT or CMR with invasive coronary angiography as the reference standard; the remaining papers referred to echocardiography (n=38), established CMR, CCT or nuclear techniques (n=58), surgical findings, biopsy or histological analysis (n=15), blood tests (n=5) or a combination of these investigations (n=17). 8 studies (2.7%) included quantitative prognostic data.

Adjudication quality of adherence to STARD

98.5% of all STARD criteria were evaluated in agreement between the reviewers. Reviewer 4 resolved disagreements on 116 of the 7500 (1.5%) assessed items. This high rate of agreement is reflected in an unweighted kappa-value of 0.96 (95% confidence interval 0.96 to 0.97).

General reporting quality in non-invasive cardiovascular imaging

The global adherence to STARD for the 300 studies was 70.6 ± 14.1% and Table 1 shows adherence for each STARD criteria separately. STARD criteria were considered 'Not Applicable' in 45 of 7500 (0.6%) assessed items. The journal impact factor in the year of publication and the citation index for all non-invasive cardiovascular imaging studies were 2.8 (1.8; 4.8) and 5.0 (1.0; 11.0), respectively.

Reporting quality according to imaging modality and impact of STARD initiative

The adherence (%) to STARD criteria across the six groups - CCT, SPECT and CMR before and after STARD introduction in 2003 - is presented for each criterion separately (Table 2) and for the combined criteria (Table 3).

The reporting quality increased from 65.3% to 74.1% (p=0.003) for CMR studies and from 61.6% to 79.0% (p<0.001) for CCT studies following the introduction of STARD in 2003. The SPECT studies, however, did not show any significant change in reporting quality: 71.9% before and 71.5% after STARD (p=0.92). Before the introduction of STARD, CCT studies had significantly poorer reporting standards compared to SPECT studies (p=0.001). After the introduction of STARD, CCT reporting standards were significantly higher than those of the SPECT studies (p=0.008). All other group comparisons did not show any significant difference in reporting quality (p>0.05 for all).

Reporting quality according to journal's author instructions, impact factors and citation indices

Since 2003, papers from journals (13/150=8.7%) that advised authors to refer to the STARD guidelines demonstrated a trend of higher reporting quality when compared to those that did not (81.2 ±9.3% vs. 74.3 ± 12.7%, respectively; p=0.06). However, after adjusting for the potential confounder impact factor and year of publication, this difference attenuated (p=0.72). ~~However, journals that advised authors to comply with the STARD criteria did not have a greater increase in adherence to STARD compared with those journals that did not after adjusting for impact factor and year of publication (p=0.72).~~

Formatted: Font: Times New Roman, 10 pt

The impact factor of journals that have adopted STARD was significantly higher than those that have not (5.3 (3.7; 5.7) vs. 2.8 (2.1; 4.0), respectively; p=0.03). In journals whose impact factor was equal to or above the median, reporting standards were significantly higher than in journals with lower impact factors (Figure 3 – p<0.0001). We further investigated whether the impact factor benefited for journals that recommended adherence to STARD criteria. To this end, the null hypothesis that the slope or rate of change for the impact factor after publication of the STARD publication is the same for journals that recommended adherence to STARD criteria compared to those that did not could not be rejected, after adjusting for year of publication (p=0.15).

The citation index for the two years following publication was similar between studies published before and after the STARD initiative (7.5 (5.0; 23.0) vs. 4.0 (2.0; 12.0) respectively; p=0.14).

Impact of the STARD initiative on reporting quality when controlling for confounders

Our multivariable linear regression model allowed for a change in slope by introducing a knot in year 2003 (STARD publication), which demonstrates the beneficial effect of the STARD criteria on the reporting standards of diagnostic accuracy studies. After controlling for the potential confounder impact factor, reporting quality remained unchanged in the 8 years before the introduction of STARD (estimate -0.0187, p=0.97). However, during the 8 years after the introduction of STARD, the model suggests an increase in adherence to STARD criteria by 1.461% (sum of estimates -0.018796 +1.479, p=0.04) for each calendar year (Figure 2).

DISCUSSION

The important findings of this study are firstly that the reporting quality of studies investigating the diagnostic accuracy of CCT, SPECT and CMR techniques is at most satisfactory. Furthermore, since the publication of the STARD statement in 2003, reporting standards have significantly improved in studies of CCT and CMR but not SPECT. Our assessment also shows that higher reporting quality is more strongly associated with a journal's impact factor than with the journal mentioning the STARD criteria in the authors' instructions, and that reporting quality does not correlate with citation index.

Whilst similar reviews have been performed in fields such as Endoscopy [6] and Ophthalmology [7, 8], this is the first investigation into the standards of CCT, SPECT and CMR studies published both before and after the STARD statement. An overall average of 70.6% adherence to the STARD criteria compares favourably with findings from similar reviews of endoscopy (49%) (6), ophthalmology (50.3%) [7, 8] and gynaecology (55.1%) [9] journals.

Previous reviews on the impact of the STARD statement itself have been mixed. Whilst Smidt et al. (2006) [10] reported a significant improvement in reporting standards across a sample of 265 articles from 12 medical journals, Wilczynski [11] did not find any meaningful improvement when comparing studies published before and after 2003, nor any difference between articles from journals that had adopted the STARD statement and those that had not. In our study, after adjustment for the confounder impact factor, the reporting standards measured by adherence to STARD criteria improved by an estimated mean of 1.5% per calendar year after the publication of the STARD statement.

We use adherence to the STARD statement as the sole measure of reporting quality, although the 25 criteria are not all-encompassing; for example, they make no stipulation of minimum sample size and only require a discussion of the clinical applicability of the study findings. However, they are specifically tailored to diagnostic accuracy studies and we believe should be considered the gold standard for reporting quality in this study type. Even so, in Hirst & Altman's [12] review of 116 journals, only 19 (16.4%) referred to reporting guidelines in their online instructions for authors, and we found that only 8.7% advised reference to STARD when submitting diagnostic accuracy studies. Our study suggests that reporting quality improves with this requirement.

The finding that reporting standards have improved in CCT and CMR studies but not in those of SPECT is intriguing. This may be because, prior to STARD, CCT and CMR studies were predominantly developmental in nature and have since progressed to the validation phase. SPECT, meanwhile, is longer established and its use, together with other nuclear techniques, is declining by 3% annually in the United States [2]. One could also speculate that, as CCT and CMR are younger imaging modalities, the researchers publishing in these areas are more willing to adapt to new reporting guidelines.

Given the potentially serious consequences of poor diagnostic studies on patient management and healthcare costs, reporting standards remain a concern despite the clear improvement seen following publication of the STARD statement. The reporting of individual criteria varied from 28-98% , with seven criteria addressed in less than 50% of studies. Only 33% of studies made estimations of test reproducibility, which echoes that of similar reviews [8].This is likely attributable to the considerable resources required to perform such measurements and many authors choose to publish separate studies of reproducibility. However, the ability of a diagnostic test to deliver reproducible results is paramount, and every effort should be made to provide or reference such data where feasible. A quantification of observers' training was found in only 28% of studies; this may seem a stringent requirement but, in practice, clinicians often consider an observer's expertise when evaluating their opinion, and so such information should be made available to study readers as it informs generalizability of the study findings. Only 44% of studies mentioned adverse events; one might presume this reflects the relative safety of CCT, SPECT and CMR techniques over their invasive counterparts, but clarification of the absence of any complications is still necessary.

Strategies to further improve reporting standards may include journals not only advising reference to the STARD criteria in their author instructions, but making adherence to STARD a mandatory prerequisite to manuscript submission. In addition, the publication of systematic reviews such as this may further increase awareness of the STARD criteria and the importance of adhering to them.

Although the selection of 300 studies constitutes less than 10% of the total identified literature, our selection and randomisation process should have ensured a representative sample of diagnostic accuracy manuscripts. As we did not adjust for multiple testing, there is a chance of reporting false positive findings. However, we pre-specified the assessment of the impact of the STARD recommendations on the quality of reporting as our primary hypothesis. We were surprised by the large disparity in the number of studies identified before and after 2003, but this could be attributed to the increased volume of studies published. Having used "diagnostic accuracy" as a search term, there is potential for a selection bias. The decision to include a small number of prognostic studies (n=8) must also be justified; the STARD statement refers to diagnostic accuracy studies alone, however, we believe the criteria are equally applicable to prognostic studies investigating quantitative outcome variables – such as degree of residual mitral regurgitation – where an account of technical specifications, statistical models, time intervals, participant recruitment and demographics, dropout and outliers remains essential.

The reporting standards of diagnostic accuracy studies in the field of non-invasive cardiac imaging are satisfactory at best and have improved since the introduction of STARD. Those journals that advise authors to refer to STARD have significantly higher impact factors, and authors should be encouraged that journals of relatively high impact factors publish diagnostic accuracy studies of higher reporting quality. To further increase the adherence to the STARD criteria and thereby improve the quality of diagnostic accuracy studies, we suggest that more journals incorporate the STARD statement as a mandatory component of their submission process. By improving the transparency and completeness of study reporting, such measures may expedite the development of non-invasive imaging tests, reduce unnecessary expenditure and assist doctors in making evidence-based management decisions.

ACKNOWLEDGMENTS

This work forms part of the research themes contributing to the translational research portfolio of Barts Cardiovascular Biomedical Research Unit which is supported and funded by the National Institute for Health Research.

REFERENCES

1. Iglehart JK**.** Health insurers and medical-imaging policy--a work in progress. N Engl J Med. 2009;360(10):1030-7.
2. Smith-Bindman R, Miglioretti DL, Johnson E, Lee C, Feigelson HS, Flynn M, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA. 2012;307(22):2400-9.
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138(1):40-4.
4. 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. The Standards for Reporting of Diagnostic Accuracy Group. Croat Med J. 2003;44(5):639-50.
5. Vandenbroucke JP**.** STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUOROM, REMARK... and CONSORT: for whom does the guideline toll? J Clin Epidemiol. 2009;62(6):594-6.
6. Areia M, Soares M, Dinis-Ribeiro M**.** Quality reporting of endoscopic diagnostic studies in gastrointestinal journals: where do we stand on the use of the STARD and CONSORT statements? Endoscopy. 2010;42(2):138-47.
7. Johnson ZK, Siddiqui MA, Azuara-Blanco A**.** The quality of reporting of diagnostic accuracy studies of optical coherence tomography in glaucoma. Ophthalmology. 2007;114(9):1607-12.
8. Siddiqui MA, Azuara-Blanco A, Burr J**.** The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. Br J Ophthalmol. 2005;89(3):261-5.
9. Selman TJ, Khan KS, Mann CH**.** An evidence-based approach to test accuracy studies in gynecologic oncology: the 'STARD' checklist. Gynecol Oncol. 2005;96(3):575-8.
10. 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(5):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(3):817-23.

12. Hirst A, Altman DG**.** Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. PLoS One. 2012;7(4):e35621.

Table 1.

STARD CRITERIA	Number of studies that met criteria	Number of studies that did not meet criteria	Number of studies where criteria not applicable	ADHERENCE (%) (*n=300*)
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.	295	5	0	98
State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	**290**	**10**	**0**	**97**
Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	286	14	0	95
Discuss the clinical applicability of the study findings.	**286**	**14**	**0**	**95**
Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the (evaluated) index tests or the (golden) reference standard?	285	15	0	95
Describe the reference standard and its rationale.	**272**	**28**	**0**	**91**
Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	259	41	0	86
Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	**256**	**43**	**1**	**85**
Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	256	44	0	85

Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	**249**	**51**	**0**	**83**
Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	247	53	0	82
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.	**237**	**61**	**2**	**79**
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.	230	68	2	77
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.	**226**	**73**	**1**	75
Report time interval from the index tests to the reference standard, and any treatment administered between.	222	77	1	74
Identify the article as a study of diagnostic accuracy(recommend MeSH heading 'sensitivity and specificity').	**210**	**90**	**0**	**70**
Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers).	208	91	1	69
Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	**166**	**133**	**1**	55

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).	135	150	15	45
Report any adverse events from performing the index tests or the reference standard.	**131**	**169**	**0**	**44**
Report when study was done, including beginning and ending dates of recruitment.	122	178	0	41
Report how indeterminate results, missing responses and outliers of the index tests were handled.	**109**	**190**	**1**	**36**
Describe methods for calculating test reproducibility, if done.	99	190	11	33
Report estimates of test reproducibility, if done.	**98**	**193**	**9**	**33**
Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.	85	215	0	28

Mean adherence (%) to individual STARD criteria

Table 2.

CRITERIA		**GROUP:**	**CMR before STARD**	**CMR after STARD**	***p value***	**CCT before STARD**	**CCT after STARD**	***p value***	**SPECT before STARD**	**SPECT after STARD**	***p value***
TITLE	**1**	**Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').**	74	52	**0.~~0250~~ 3**	52	90	**0.0001**	78	74	*0.82*
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	100	94	*0.24*	94	100	*0.24*	94	98	*0.62*
METHODS											
Participants	**3**	**Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.**	82	98	**0.~~0150~~ 2**	72	88	0.07	88	70	**0.~~0290~~ 5**
	4	Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the (evaluated) index tests or the (golden) reference standard?	88	96	*0.06*	94	98	*0.99*	98	96	*0.99*
	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.**	70	78	*0.49*	78	84	*0.39*	76	88	*0.19*
	6	Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	86	98	*0.06*	74	94	**0.~~0076~~ 008**	82	78	*0.8*
Test methods	**7**	**Describe the reference standard and its rationale.**	86	98	*0.06*	84	90	0.55	100	86	**0.~~0130~~**

												1
	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.	98	100	*0.99*	98	100	*0.99*	100	94	*0.24*
	9	**Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.**	84	100	**0.004**	96	98	*0.99*	100	94	*0.24*
	10	Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.	8	66	**0.0001**	22	26	0.81	40	8	**0.0003**
	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.**	64	74	*0.39*	70	90	**0.~~0150~~ 2**	80	80	*1*
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	72	92	**0.01~~1~~**	66	90	**0.~~0046~~ 005**	82	92	*0.23*
	13	**Describe methods for calculating test reproducibility, if done.**	30	34	*0.83*	14	46	**0.0007**	26	48	**0.~~0250~~ 3**
RESULTS												
Participants	14	Report when study was done, including beginning and ending dates of recruitment.	30	34	*0.83*	40	66	**0.~~0160~~ 2**	32	42	*0.4*
	15	**Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers).**	48	94	**0.0001**	52	74	**0.~~0250~~ 3**	78	70	*0.49*

	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).	44	66	**0.~~0250~~ 3**	44	60	*0.16*	26	30	*0.82*
Test results	**17**	**Report time interval from the index tests to the reference standard, and any treatment administered between.**	70	62	*0.53*	64	84	**0.~~0250~~ 3**	88	76	*0.19*
	18	Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	72	94	**0.~~0042~~ 004**	80	84	*0.79*	92	90	*0.99*
	19	**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.**	64	62	*0.99*	62	90	**0.~~0014~~ 001**	80	94	*0.07*
	20	Report any adverse events from performing the index tests or the reference standard.	58	36	**0.~~0250~~ 3**	38	68	**0.~~0034~~ 003**	34	28	*0.67*
Estimates	**21**	**Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).**	92	80	*0.14*	70	80	*0.36*	98	96	*0.99*
	22	Report how indeterminate results, missing responses and outliers of the index tests were handled.	52	42	*0.42*	30	50	*0.~~065~~0 7*	24	20	*0.81*
	23	**Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.**	30	50	*0.06*	40	58	*0.1*	66	86	**0.~~034~~0 3**

	24	Report estimates of test reproducibility, if done.	22	38	*0.12*	18	50	**0.~~0011~~ 001**	24	44	*0.~~0560~~ 6*
DISCUSSION	**25**	**Discuss the clinical applicability of the study findings.**	94	96	*0.99*	92	98	*0.36*	96	96	*1*

Adherence (%) to STARD criteria per imaging modality and timing of publication in relation to STARD statement (p values <0.05 in **bold***)*

Table 3.

		Before STARD	***After STARD***	*P value*
Adherence to STARD (%)	*CMR*	*65.3 ± 17.3*	*74.1 ± 10.1*	***0.003***
	CCT	*61.6 ± 13.0*	*79.0 ± 14.3*	***0.0001***
	SPECT	*71.7 ± 9.9*	*71.5 ± 11.8*	*0.92*
Impact factor at time of publication	*CMR*	*2.4 (1.4; 5.6)*	*3.7 (2.7; 5.9)*	***0.01***
	CCT	*2.4 (1.9; 5.8)*	*2.8 (2.2; 3.7)*	*0.92*
	SPECT	*2.4 (1.8; 4.5)*	*2.7 (1.3; 3.4)*	*0.83*
Citation index for 2 years after publication	*CMR*	*2.0 (0; 8.0)*	*5.0 (3.0; 23.0)*	***0.004***
	CCT	*9.0 (4.0; 26.0)*	*6.5 (4.0; 28.0)*	*0.8*
	SPECT	*3.0 (1.0; 6.0)*	*3.0 (2.0; 6.0)*	*0.83*

Adherence (%) to STARD, journal impact factor in for the year of publication and number of citations for the two years after publication per imaging modality and timing of publication with regards to STARD statement publication. P values for adherence are based on one-way ANOVA, and for impact factor and citation index on Kruskal-Wallis test;

FIGURE LEGENDS

Figure 1: Flow diagram of study selection process

Figure 2: Reporting quality of diagnostic accuracy studies before and after the introduction of STARD criteria in 2003. The left hand panel shows the mean +/- standard deviation of adherence (%) to STARD criteria in the 300 studies assessed. The right hand panel shows the predicted adherence (%) to STARD criteria applying the multivariable linear regression model which adjusts for impact factor of the journal.

Figure 3: % adherence to STARD for studies from journals of above and below median impact factor. Journals of above median impact factor published studies with significantly higher mean adherence to STARD (74.1% ±12.2 vs 66.7% ±13.9) when compared to those of below median impact factor ($p<0.0001$)

Reporting standards in cardiac MRI, CT and SPECT diagnostic accuracy studies: Analysis of the impact of STARD criteria

Edd N Maclean[1], Ian S Stone[1], Felix Ceelen[2], Xabier Garcia-Albeniz[3], Wieland H Sommer[2], Steffen E Petersen*[1]

Affiliations: [1]Advanced Cardiovascular Imaging, NIHR Cardiovascular Biomedical Research Unit at Barts, William Harvey Research Institute, Queen Mary University of London, UK; [2] Department of Clinical Radiology, University of Munich, Grosshadern Campus, Munich, Germany; [3] Department of Epidemiology, Harvard School of Public Health, USA.

***Corresponding author: Steffen E. Petersen, MD DPhil FRCP FESC FACC**

Professor of Cardiovascular Medicine,

Honorary Consultant Cardiologist,

Centre Lead for Advanced Cardiovascular Imaging,

William Harvey Research Institute,

NIHR Cardiovascular Biomedical Research Unit at Barts,

The London Chest Hospital,

Bonner Road,

London,

E2 9JX,

UK

Email: s.e.petersen@qmul.ac.uk

Words: 2982

ABSTRACT

Aims: Diagnostic accuracy studies determine the clinical value of non-invasive cardiac imaging tests. The 'Standards for the reporting of diagnostic accuracy studies' (STARD) were published in 2003 to improve the quality of study reporting. We aimed to assess the reporting quality of cardiac computed tomography (CCT), single positron emission computed tomography (SPECT) and cardiac magnetic resonance (CMR) diagnostic accuracy studies; to evaluate the impact of STARD; and to investigate the relationships between reporting quality, journal impact factor and study citation index.

Methods and Results: We randomly generated 6 groups of 50 diagnostic accuracy studies: 'CMR 1995-2002', 'CMR 2004-2011', 'CCT 1995-2002', 'CCT 2004-2011', 'SPECT 1995-2002', and 'SPECT 2004-2011'. The 300 studies were double-read by 2 blinded reviewers and reporting quality determined by % adherence to the 25 STARD criteria. Reporting quality increased from 65.3% before STARD to 74.1% after ($p=0.003$) in CMR studies and from 61.6% to 79.0% ($p<0.001$) in CCT studies. SPECT studies showed no significant change: 71.9% before and 71.5% after STARD ($p=0.92$). Journals advising authors to refer to STARD had significantly higher impact factors than those that did not ($p=0.03$), and journals with above-median impact factors published studies of significantly higher reporting quality ($p<0.001$). Since STARD, citation index has not significantly increased ($p=0.14$) but, after adjustment for impact factor, reporting quality continues to increase by approximately 1.5% each year.

Conclusion: Reporting standards for diagnostic accuracy studies of non-invasive cardiac imaging are at most satisfactory and have improved since the introduction of STARD. Adherence to STARD should be mandatory for authors of diagnostic accuracy studies.

KEYWORDS

Diagnostic accuracy, STARD, reporting quality

ABBREVIATIONS LIST

CMR, cardiac magnetic resonance;
CCT, cardiac computed tomography;
SPECT, single positron emission computed tomography;
STARD, Standards for the reporting of diagnostic accuracy studies;

INTRODUCTION

Advanced non-invasive cardiovascular imaging modalities, such as Cardiovascular Magnetic Resonance (CMR) imaging, Cardiovascular Computed Tomography (CCT) and Single Positron Emission Computed Tomography (SPECT) are increasingly requested clinically. Between 2000 and 2006, Medicare expenditure on medical imaging increased annually by 17% and, since 1996, mean per capita radiation dose has doubled, highlighting the need to avoid unnecessary use of these expensive technologies [1, 2].

Diagnostic accuracy is an important consideration in determining the cost-effectiveness of an imaging test, but often varies amongst different publications. This may reflect the dependence of the results on factors such as study design, patient population and technical considerations as well as random variability. Insufficient reporting may not allow assessment of the internal and external validity of the study findings. Furthermore, over-optimistic diagnostic accuracy results can lead to the premature dissemination of imaging tests and consequently to doctors making incorrect management decisions, contributing to the significant rise in health care costs.

In 2003, the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) group published a set of 25 criteria with the objective of improving the reporting quality of diagnostic accuracy studies [3-5]. These criteria allow the reader to identify the potential for bias in the study (internal validity) and to evaluate whether the results of the studies can be generalized to a wider population (external validity). To date, over 200 journals advise authors to refer to STARD when submitting manuscripts (http://www.stard-statement.org/).

Given the importance of high quality diagnostic test reporting in cardiac patients and the lack of data on adherence to the STARD criteria in this field, the aim of this study was to assess the impact of STARD by comparing the reporting quality of CCT, SPECT and CMR studies published in the eight years before STARD (1995-2002) with those published in the eight years after (2004-2011).

METHODS

Literature Search

A literature search of the Ovid Medline and EMBASE databases was performed on September 1st 2011. We searched for CCT, SPECT and CMR studies of diagnostic accuracy published before (1995-2002 inclusive) and after (2004-2011 inclusive) the introduction of the STARD statement in 2003. The MeSH terms, corresponding number of identified studies and study groupings are shown in Figure 1.

Selection Criteria

We included studies that examined the performance of CCT, SPECT or CMR investigations in relation to a reference standard. Animal models, reviews, meta-analyses, and studies comparing more than one of the CCT, SPECT or CMR modalities against a reference standard were excluded. The identified studies were assigned to one of six groups: CCT, SPECT or CMR diagnostic accuracy studies published before and after the introduction of STARD in 2003, respectively. In each group, eligible studies were numbered and 50 papers selected for analysis using a random number generator (Microsoft Excel 2010, Microsoft). Data extraction was performed from each of the selected abstracts. Studies that did not meet the inclusion criteria during data extraction were excluded and replaced by studies using the random number generator until 50 studies were identified for each of the six groups.

Scoring

The 300 included studies were blinded to date, authorship, references and journal of publication, and read against the criteria of the STARD checklist. Reviewer 1 (E.M.) read all 300 studies; Reviewer 2 (I.S.) read a random selection of 100 studies, and Reviewer 3 (F.C.) read the remaining 200 studies; both reviewers were blinded to the findings of Reviewer 1. Reviewer 4 (S.P.) resolved any disputed decisions. All four reviewers were provided with a document explaining the STARD statement and its rationale, and were instructed to refer to the STARD statement website (www.stard-statement.org) if further clarification on the criteria was required. A total of 7500 STARD items were evaluated within the 300 manuscripts. For each STARD criterion, reviewers assigned ‘Yes’ if the manuscript addressed

the item appropriately and 'No' if it did not. If a criterion was considered not applicable to the study, such as in retrospective studies where participant dropout does not occur, the abbreviation 'NA' was used.

Impact Factor and Citation Index

Impact factor in the year of study publication was sourced for each journal from the Thomson Reuters Web of Science database. Study Citation Index was calculated by counting the total number of citations in the two years following study publication according to Web of Science data.

Statistics

Reporting quality was assessed in all studies by calculating the percentage adherence to the STARD criteria by dividing the number of agreements with STARD criteria per study by the number of possible agreements (25 criteria minus number of criteria considered not applicable for specific study). We used the following formula:

$$\% \text{ Adherence} = \frac{\text{Number of 'Yes'}}{25 - \text{Number of 'NA'}} \times 100$$

Data were examined for normality (median and mean comparison, skewness, kurtosis, the Shapiro Wilks test and normal probability plots). Normally distributed data are presented as mean ± standard deviation and non-normally distributed data as median (interquartile range). We performed independent t-tests or ANOVA for normally distributed data and the Wilcoxon rank-sum test or Kruskal-Wallis test for independent samples that were not normally distributed. Chi-square tests were used to assess for differences in adherence to all individual STARD criteria.

A linear regression model was built with adherence (%) to STARD criteria as the outcome and the timing of publication with regards to the advent of the STARD criteria in 2003 (before or after STARD) as a binary exposure and potential confounders (impact factor, citation index). A spline with a knot in the year of STARD publication was introduced to allow for a change in the slope. In all cases the significance level was set at $p \leq 0.05$ (two-sided). No adjustment for multiple testing was performed for pre-specified sub-analyses. All statistical analyses were performed using SAS software (Version 9.3; SAS Institute Inc., Cary, NC, US).

RESULTS

Study Selection

As shown in Figure 1, 37 of the initial 300 studies met exclusion criteria and were replaced; 19 were animal studies, 8 were meta-analyses or reviews, 6 studies compared more than one imaging modality to a reference standard, and 4 studies did not examine diagnostic accuracy. Of the included studies, 167 evaluated the diagnostic accuracy of CCT, SPECT or CMR with invasive coronary angiography as the reference standard; the remaining papers referred to echocardiography (n=38), established CMR, CCT or nuclear techniques (n=58), surgical findings, biopsy or histological analysis (n=15), blood tests (n=5) or a combination of these investigations (n=17). 8 studies (2.7%) included quantitative prognostic data.

Adjudication quality of adherence to STARD

98.5% of all STARD criteria were evaluated in agreement between the reviewers. Reviewer 4 resolved disagreements on 116 of the 7500 (1.5%) assessed items. This high rate of agreement is reflected in an unweighted kappa-value of 0.96 (95% confidence interval 0.96 to 0.97).

General reporting quality in non-invasive cardiovascular imaging

The global adherence to STARD for the 300 studies was 70.6 ± 14.1% and Table 1 shows adherence for each STARD criteria separately. STARD criteria were considered 'Not Applicable' in 45 of 7500 (0.6%) assessed items. The journal impact factor in the year of publication and the citation index for all non-invasive cardiovascular imaging studies were 2.8 (1.8; 4.8) and 5.0 (1.0; 11.0), respectively.

Reporting quality according to imaging modality and impact of STARD initiative

The adherence (%) to STARD criteria across the six groups - CCT, SPECT and CMR before and after STARD introduction in 2003 - is presented for each criterion separately (Table 2) and for the combined criteria (Table 3).

The reporting quality increased from 65.3% to 74.1% (p=0.003) for CMR studies and from 61.6% to 79.0% (p<0.001) for CCT studies following the introduction of STARD in 2003. The SPECT studies, however, did not show any significant change in reporting quality: 71.9% before and 71.5% after STARD (p=0.92). Before the introduction of STARD, CCT studies had significantly poorer reporting standards compared to SPECT studies (p=0.001). After the introduction of STARD, CCT reporting standards were significantly higher than those of the SPECT studies (p=0.008). All other group comparisons did not show any significant difference in reporting quality (p>0.05 for all).

Reporting quality according to journal's author instructions, impact factors and citation indices

Since 2003, papers from journals (13/150=8.7%) that advised authors to refer to the STARD guidelines demonstrated a trend of higher reporting quality when compared to those that did not (81.2 ±9.3% vs. 74.3 ± 12.7%, respectively; p=0.06). However, after adjusting for the potential confounder impact factor and year of publication, this difference attenuated (p=0.72).

The impact factor of journals that have adopted STARD was significantly higher than those that have not (5.3 (3.7; 5.7) vs. 2.8 (2.1; 4.0), respectively; p=0.03). In journals whose impact factor was equal to or above the median, reporting standards were significantly higher than in journals with lower impact factors (Figure 3 – p<0.0001). We further investigated whether the impact factor benefited for journals that recommended adherence to STARD criteria. To this end, the null hypothesis that the slope or rate of change for the impact factor after publication of the STARD publication is the same for journals that recommended adherence to STARD criteria compared to those that did not could not be rejected, after adjusting for year of publication (p=0.15).

The citation index for the two years following publication was similar between studies published before and after the STARD initiative (7.5 (5.0; 23.0) vs. 4.0 (2.0; 12.0) respectively; p=0.14).

Impact of the STARD initiative on reporting quality when controlling for confounders

Our multivariable linear regression model allowed for a change in slope by introducing a knot in year 2003 (STARD publication), which demonstrates the beneficial effect of the STARD criteria on the

reporting standards of diagnostic accuracy studies. After controlling for the potential confounder impact factor, reporting quality remained unchanged in the 8 years before the introduction of STARD (estimate -0.018, p=0.97). However, during the 8 years after the introduction of STARD, the model suggests an increase in adherence to STARD criteria by 1.461% (sum of estimates -0.018 +1.479, p=0.04) for each calendar year (Figure 2).

DISCUSSION

The important findings of this study are firstly that the reporting quality of studies investigating the diagnostic accuracy of CCT, SPECT and CMR techniques is at most satisfactory. Furthermore, since the publication of the STARD statement in 2003, reporting standards have significantly improved in studies of CCT and CMR but not SPECT. Our assessment also shows that higher reporting quality is more strongly associated with a journal's impact factor than with the journal mentioning the STARD criteria in the authors' instructions, and that reporting quality does not correlate with citation index.

Whilst similar reviews have been performed in fields such as Endoscopy [6] and Ophthalmology [7, 8], this is the first investigation into the standards of CCT, SPECT and CMR studies published both before and after the STARD statement. An overall average of 70.6% adherence to the STARD criteria compares favourably with findings from similar reviews of endoscopy (49%) (6), ophthalmology (50.3%) [7, 8] and gynaecology (55.1%) [9] journals.

Previous reviews on the impact of the STARD statement itself have been mixed. Whilst Smidt et al. (2006) [10] reported a significant improvement in reporting standards across a sample of 265 articles from 12 medical journals, Wilczynski [11] did not find any meaningful improvement when comparing studies published before and after 2003, nor any difference between articles from journals that had adopted the STARD statement and those that had not. In our study, after adjustment for the confounder impact factor, the reporting standards measured by adherence to STARD criteria improved by an estimated mean of 1.5% per calendar year after the publication of the STARD statement.

We use adherence to the STARD statement as the sole measure of reporting quality, although the 25 criteria are not all-encompassing; for example, they make no stipulation of minimum sample size and only require a discussion of the clinical applicability of the study findings. However, they are

specifically tailored to diagnostic accuracy studies and we believe should be considered the gold standard for reporting quality in this study type. Even so, in Hirst & Altman's [12] review of 116 journals, only 19 (16.4%) referred to reporting guidelines in their online instructions for authors, and we found that only 8.7% advised reference to STARD when submitting diagnostic accuracy studies. Our study suggests that reporting quality improves with this requirement.

The finding that reporting standards have improved in CCT and CMR studies but not in those of SPECT is intriguing. This may be because, prior to STARD, CCT and CMR studies were predominantly developmental in nature and have since progressed to the validation phase. SPECT, meanwhile, is longer established and its use, together with other nuclear techniques, is declining by 3% annually in the United States [2]. One could also speculate that, as CCT and CMR are younger imaging modalities, the researchers publishing in these areas are more willing to adapt to new reporting guidelines.

Given the potentially serious consequences of poor diagnostic studies on patient management and healthcare costs, reporting standards remain a concern despite the clear improvement seen following publication of the STARD statement. The reporting of individual criteria varied from 28-98% , with seven criteria addressed in less than 50% of studies. Only 33% of studies made estimations of test reproducibility, which echoes that of similar reviews [8].This is likely attributable to the considerable resources required to perform such measurements and many authors choose to publish separate studies of reproducibility. However, the ability of a diagnostic test to deliver reproducible results is paramount, and every effort should be made to provide or reference such data where feasible. A quantification of observers' training was found in only 28% of studies; this may seem a stringent requirement but, in practice, clinicians often consider an observer's expertise when evaluating their opinion, and so such information should be made available to study readers as it informs generalizability of the study findings. Only 44% of studies mentioned adverse events; one might presume this reflects the relative safety of CCT, SPECT and CMR techniques over their invasive counterparts, but clarification of the absence of any complications is still necessary.

Strategies to further improve reporting standards may include journals not only advising reference to the STARD criteria in their author instructions, but making adherence to STARD a mandatory

prerequisite to manuscript submission. In addition, the publication of systematic reviews such as this may further increase awareness of the STARD criteria and the importance of adhering to them.

Although the selection of 300 studies constitutes less than 10% of the total identified literature, our selection and randomisation process should have ensured a representative sample of diagnostic accuracy manuscripts. As we did not adjust for multiple testing, there is a chance of reporting false positive findings. However, we pre-specified the assessment of the impact of the STARD recommendations on the quality of reporting as our primary hypothesis. We were surprised by the large disparity in the number of studies identified before and after 2003, but this could be attributed to the increased volume of studies published. Having used "diagnostic accuracy" as a search term, there is potential for a selection bias. The decision to include a small number of prognostic studies (n=8) must also be justified; the STARD statement refers to diagnostic accuracy studies alone, however, we believe the criteria are equally applicable to prognostic studies investigating quantitative outcome variables – such as degree of residual mitral regurgitation – where an account of technical specifications, statistical models, time intervals, participant recruitment and demographics, dropout and outliers remains essential.

The reporting standards of diagnostic accuracy studies in the field of non-invasive cardiac imaging are satisfactory at best and have improved since the introduction of STARD. Those journals that advise authors to refer to STARD have significantly higher impact factors, and authors should be encouraged that journals of relatively high impact factors publish diagnostic accuracy studies of higher reporting quality. To further increase the adherence to the STARD criteria and thereby improve the quality of diagnostic accuracy studies, we suggest that more journals incorporate the STARD statement as a mandatory component of their submission process. By improving the transparency and completeness of study reporting, such measures may expedite the development of non-invasive imaging tests, reduce unnecessary expenditure and assist doctors in making evidence-based management decisions.

ACKNOWLEDGMENTS

This work forms part of the research themes contributing to the translational research portfolio of Barts Cardiovascular Biomedical Research Unit which is supported and funded by the National Institute for Health Research.

REFERENCES

1. Iglehart JK**.** Health insurers and medical-imaging policy--a work in progress. N Engl J Med. 2009;360(10):1030-7.
2. Smith-Bindman R, Miglioretti DL, Johnson E, Lee C, Feigelson HS, Flynn M, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA. 2012;307(22):2400-9.
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138(1):40-4.
4. 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. The Standards for Reporting of Diagnostic Accuracy Group. Croat Med J. 2003;44(5):639-50.
5. Vandenbroucke JP**.** STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUOROM, REMARK... and CONSORT: for whom does the guideline toll? J Clin Epidemiol. 2009;62(6):594-6.
6. Areia M, Soares M, Dinis-Ribeiro M**.** Quality reporting of endoscopic diagnostic studies in gastrointestinal journals: where do we stand on the use of the STARD and CONSORT statements? Endoscopy. 2010;42(2):138-47.
7. Johnson ZK, Siddiqui MA, Azuara-Blanco A**.** The quality of reporting of diagnostic accuracy studies of optical coherence tomography in glaucoma. Ophthalmology. 2007;114(9):1607-12.
8. Siddiqui MA, Azuara-Blanco A, Burr J**.** The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. Br J Ophthalmol. 2005;89(3):261-5.
9. Selman TJ, Khan KS, Mann CH**.** An evidence-based approach to test accuracy studies in gynecologic oncology: the 'STARD' checklist. Gynecol Oncol. 2005;96(3):575-8.
10. 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(5):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(3):817-23.

12. Hirst A, Altman DG**.** Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. PLoS One. 2012;7(4):e35621.

Table 1.

STARD CRITERIA	**Number of studies that met criteria**	**Number of studies that did not meet criteria**	**Number of studies where criteria not applicable**	**ADHERENCE (%) (*n=300*)**
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.	295	5	0	98
State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	**290**	**10**	**0**	**97**
Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	286	14	0	95
Discuss the clinical applicability of the study findings.	**286**	**14**	**0**	**95**
Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the (evaluated) index tests or the (golden) reference standard?	285	15	0	95
Describe the reference standard and its rationale.	**272**	**28**	**0**	**91**
Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	259	41	0	86
Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	**256**	**43**	**1**	**85**
Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	256	44	0	85

Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	**249**	**51**	**0**	**83**
Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	247	53	0	82
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.	**237**	**61**	**2**	**79**
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.	230	68	2	77
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.	**226**	**73**	**1**	**75**
Report time interval from the index tests to the reference standard, and any treatment administered between.	222	77	1	74
Identify the article as a study of diagnostic accuracy(recommend MeSH heading 'sensitivity and specificity').	**210**	**90**	**0**	**70**
Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers).	208	91	1	69
Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	**166**	**133**	**1**	**55**

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).	135	150	15	45
Report any adverse events from performing the index tests or the reference standard.	**131**	**169**	**0**	**44**
Report when study was done, including beginning and ending dates of recruitment.	122	178	0	41
Report how indeterminate results, missing responses and outliers of the index tests were handled.	**109**	**190**	**1**	**36**
Describe methods for calculating test reproducibility, if done.	99	190	11	33
Report estimates of test reproducibility, if done.	**98**	**193**	**9**	**33**
Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.	85	215	0	28

Mean adherence (%) to individual STARD criteria

Table 2.

CRITERIA		**GROUP:**	**CMR before STARD**	**CMR after STARD**	***p value***	**CCT before STARD**	**CCT after STARD**	***p value***	**SPECT before STARD**	**SPECT after STARD**	***p value***
TITLE	**1**	**Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').**	74	52	**0.03**	52	90	**0.0001**	78	74	*0.82*
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	100	94	*0.24*	94	100	*0.24*	94	98	*0.62*
METHODS											
Participants	**3**	**Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.**	82	98	**0.02**	72	88	0.07	88	70	**0.05**
	4	Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the (evaluated) index tests or the (golden) reference standard?	88	96	*0.06*	94	98	*0.99*	98	96	*0.99*
	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.**	70	78	*0.49*	78	84	*0.39*	76	88	*0.19*
	6	Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	86	98	*0.06*	74	94	**0.008**	82	78	*0.8*
Test methods	**7**	**Describe the reference standard and its rationale.**	86	98	*0.06*	84	90	0.55	100	86	**0.01**
	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.	98	100		98	100		100	94	

					0.99			*0.99*			*0.24*	
	9	**Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.**	84	100	**0.004**	96	98	*0.99*	100	94	*0.24*	
	10	Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.	8	66	**0.0001**	22	26	0.81	40	8	**0.0003**	
	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.**	64	74	*0.39*	70	90	**0.02**	80	80	*1*	
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	72	92	**0.01**	66	90	**0.005**	82	92	*0.23*	
	13	**Describe methods for calculating test reproducibility, if done.**	30	34	*0.83*	14	46	**0.0007**	26	48	**0.03**	
RESULTS												
Participants	14	Report when study was done, including beginning and ending dates of recruitment.	30	34	*0.83*	40	66	**0.02**	32	42	*0.4*	
	15	**Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers).**	48	94	**0.0001**	52	74	**0.03**	78	70	*0.49*	
	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).	44	66	**0.03**	44	60	*0.16*	26	30	*0.82*	
Test results	**17**	**Report time interval from the index tests to the reference standard, and any treatment administered between.**	70	62	*0.53*	64	84	**0.03**	88	76	*0.19*	

	18	Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	72	94	**0.004**	80	84	*0.79*	92	90	*0.99*
	19	**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.**	64	62	*0.99*	62	90	**0.001**	80	94	*0.07*
	20	Report any adverse events from performing the index tests or the reference standard.	58	36	**0.03**	38	68	**0.003**	34	28	*0.67*
Estimates	**21**	**Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).**	92	80	*0.14*	70	80	*0.36*	98	96	*0.99*
	22	Report how indeterminate results, missing responses and outliers of the index tests were handled.	52	42	*0.42*	30	50	*0.07*	24	20	*0.81*
	23	**Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.**	30	50	*0.06*	40	58	*0.1*	66	86	**0.03**
	24	Report estimates of test reproducibility, if done.	22	38	*0.12*	18	50	**0.001**	24	44	*0.06*
DISCUSSION	**25**	**Discuss the clinical applicability of the study findings.**	94	96	*0.99*	92	98	*0.36*	96	96	*1*

Adherence (%) to STARD criteria per imaging modality and timing of publication in relation to STARD statement (p values <0.05 in **bold***)*

Table 3.

		Before STARD	***After STARD***	*P value*
Adherence to STARD (%)	*CMR*	*65.3 ± 17.3*	*74.1 ± 10.1*	***0.003***
	CCT	*61.6 ± 13.0*	*79.0 ± 14.3*	***0.0001***
	SPECT	*71.7 ± 9.9*	*71.5 ± 11.8*	*0.92*
Impact factor at time of publication	*CMR*	*2.4 (1.4; 5.6)*	*3.7 (2.7; 5.9)*	***0.01***
	CCT	*2.4 (1.9; 5.8)*	*2.8 (2.2; 3.7)*	*0.92*
	SPECT	*2.4 (1.8; 4.5)*	*2.7 (1.3; 3.4)*	*0.83*
Citation index for 2 years after publication	*CMR*	*2.0 (0; 8.0)*	*5.0 (3.0; 23.0)*	***0.004***
	CCT	*9.0 (4.0; 26.0)*	*6.5 (4.0; 28.0)*	*0.8*
	SPECT	*3.0 (1.0; 6.0)*	*3.0 (2.0; 6.0)*	*0.83*

Adherence (%) to STARD, journal impact factor in for the year of publication and number of citations for the two years after publication per imaging modality and timing of publication with regards to STARD statement publication. P values for adherence are based on one-way ANOVA, and for impact factor and citation index on Kruskal-Wallis test;

FIGURE LEGENDS

Figure 1: Flow diagram of study selection process

Figure 2: Reporting quality of diagnostic accuracy studies before and after the introduction of STARD criteria in 2003. The left hand panel shows the mean +/- standard deviation of adherence (%) to STARD criteria in the 300 studies assessed. The right hand panel shows the predicted adherence (%) to STARD criteria applying the multivariable linear regression model which adjusts for impact factor of the journal.

Figure 3: % adherence to STARD for studies from journals of above and below median impact factor. Journals of above median impact factor published studies with significantly higher mean adherence to STARD (74.1% ±12.2 vs 66.7% ±13.9) when compared to those of below median impact factor ($p<0.0001$)

Words: 2982

Figure 1
Click here to download high resolution image

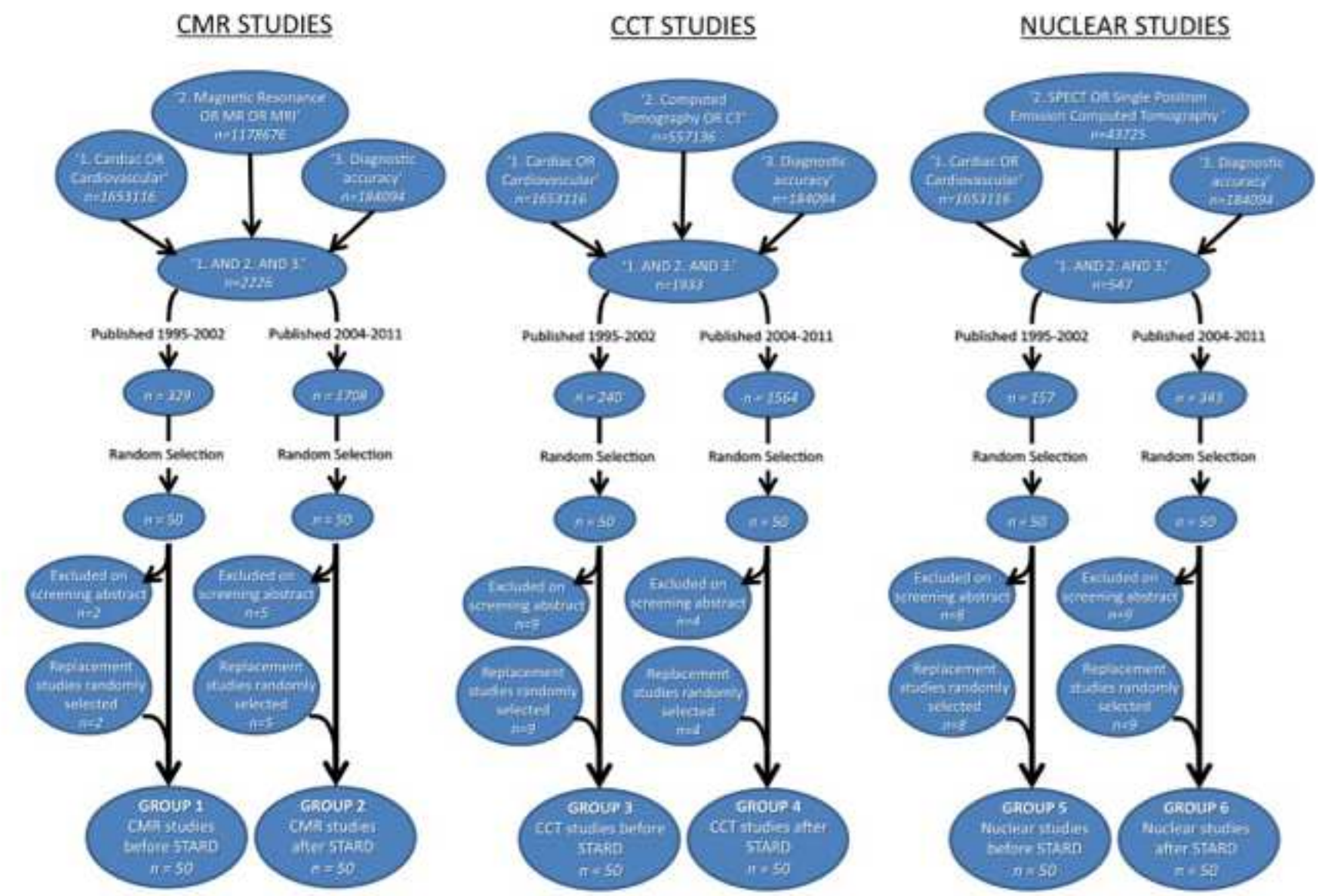

Figure 3
Click here to download high resolution image

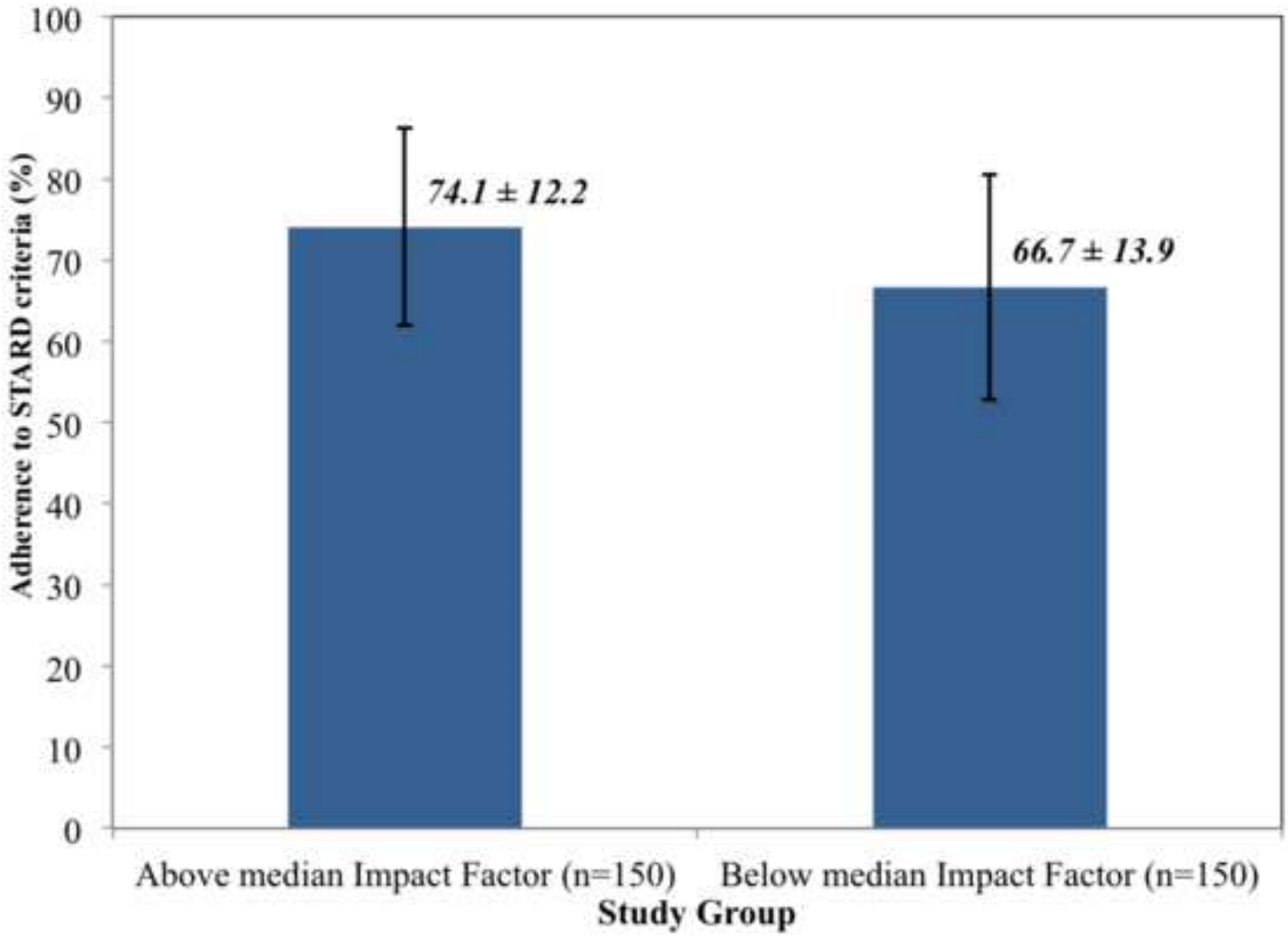

Figure 2 revised or reresubmission

Click here to download high resolution image