Published in final edited form as:
J Emerg Med. 2020 April ; 58(4): 636–646. doi:10.1016/j.jemermed.2019.09.029.

Emergency Ultrasound Literature and Adherence to Standards for Reporting of Diagnostic Accuracy Criteria

Molly Thiessen, MD[*,†], Jody A. Vogel, MD, MSC[*,†], Richard L. Byyny, MD, MSC[*,†], Emily Hopkins, MSPH[*], Jason S. Haukoos, MD, MSC[*,†,‡], John L. Kendall, MD[*,†], Stacy A. Trent, MD, MPH[*,†]

[*]Department of Emergency Medicine, Denver Health Medical Center, Denver, Colorado,

[†]Department of Emergency Medicine, University of Colorado School of Medicine, Aurora, Colorado,

[‡]Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado

Abstract

Background: Given the wide usage of Emergency point-of-care ultrasound (EUS) among emergency physicians (EPs), rigorous study surrounding its accuracy is essential. The Standards for Reporting of Diagnostic Accuracy (STARD) criteria were established to ensure robust reporting methodology for diagnostic studies. Adherence to the STARD criteria among EUS diagnostic studies has yet to be reported.

Objectives: Our objective was to evaluate a body of EUS literature shortly after STARD publication for its baseline adherence to the STARD criteria.

Methods: EUS studies in 5 emergency medicine journals from 2005–2010 were evaluated for their adherence to the STARD criteria. Manuscripts were selected for inclusion if they reported original research and described the use of one of ten diagnostic ultrasound modalities designated as “core emergency ultrasound applications” in the 2008 ACEP Ultrasound Guidelines. Literature search identified 307 studies; of these, 45 met inclusion criteria for review.

Results: The median STARD score was 15 (IQR: 12 – 17), representing 60% of the 25 total STARD criteria. The median STARD score among articles that reported diagnostic accuracy was significantly higher than those that did not report accuracy [17 IQR: 15 – 19) vs. 11 (IQR: 9 – 13), respectively; $p < 0.0001$]. Seventy one percent of articles met at least 50% of the STARD criteria (56%−84%) and 4% met greater than 80% of the STARD criteria.

Conclusions: Significant opportunities exist to improve methodological reporting of EUS research. Increased adherence to the STARD criteria among diagnostic EUS studies will improve reporting and ability to compare outcomes.

Corresponding Address: Molly Thiessen, MD, Department of Emergency Medicine, Denver Health Medical Center, 777 Bannock St., Mail Code 0108, Denver, CO 80204.

Keywords

point-of-care ultrasound; emergency medicine; diagnostic accuracy; Standards for the Reporting of Diagnostic Accuracy; STARD; research methodology

INTRODUCTION

Emergency point-of-care ultrasound (EUS) is widely used in emergency departments (EDs) (ACEP, 2008). Multiple studies have shown a significant utility of EUS as a diagnostic tool, and its use in clinical care is expanding. As with any diagnostic tool, demonstration of diagnostic accuracy prior to its widespread implementation is important. Unfortunately, methodological flaws in study design and execution may result in biased estimates of accuracy, thus limiting the applicability of such modalities in clinical care settings[1,2]. Moreover, poor diagnostic accuracy may result in inappropriate utilization of diagnostic tests by physicians[3].

In 2003, the Standards for Reporting of Diagnostic Accuracy (STARD) criteria were published with a goal of improving scientific reporting of studies that evaluate diagnostic tests. Using prior guidelines and expert opinion, STARD consists of 25 items that are largely considered requisite when publishing study results[3] (Table 1).

It remains unclear how well published EUS research adheres to STARD criteria. Thus, our objective was to evaluate a representative body of EUS literature with a hypothesis that significant variability and poor adherence with STARD criteria would exist. Thus, our objective was to evaluate a body of EUS literature and its adherance to the STARD criteria.

METHODS

This project included only aggregate non-human subjects data from previously published work. As such, no institutional review board approval was necessary.

Study Design

We performed a systematic review of emergency medicine EUS literature.

Search Strategy

We targeted five peer-reviewed, emergency medicine journals based on impact factor and readership: *Annals of Emergency Medicine*, *Academic Emergency Medicine*, *American Journal of Emergency Medicine*, *Emergency Medicine Journal*, and *Journal of Emergency Medicine*. A manual search of all of the issues for each of these journals from January 1, 2005 through December 31, 2010 was conducted by the primary author to identify any articles related to EUS. This search consisted of a review of the table of contents to identify any article with title or keywords pertaining to EUS. When the use of EUS was unclear based on the title or keywords, the abstract and article were reviewed. This manual search was then supplemented by a subsequent electronic search of the same journals over the same time period using PubMed. The search terms “emergency” and “ultrasound” or “echo” or "sono" were used, and then filtered by each journal queried. No additional articles were found.

Article Selection

Manuscripts were selected by the primary author for inclusion if they reported original research and described the use of one of the ten ultrasound modalities designated as "core emergency ultrasound applications" in the 2008 ACEP Ultrasound Guidelines for diagnostic purposes in the ED. The ten core EUS applications included: trauma, intrauterine pregnancy (IUP), abdominal aortic aneurysm (AAA), cardiac, biliary, urinary tract, deep venous thrombosis (DVT), soft-tissue/musculoskeletal (ST/MSK), thoracic, and ocular[4]. Articles specific to procedural guidance were excluded, as these are not amenable to analysis via the STARD criteria. We also excluded publications that dealt with new or novel ultrasound applications not listed in the core applications, applications that did not use EUS as a diagnostic tool, or educational studies and case reports.

Data Abstraction

Once a complete list of articles was identified, the articles were blinded for review based on the journal in which the article was published and the study site and authorship group of the study. Each article was copied and pasted into a word processing program (Microsoft Word, Microsoft Corporation, Redmond, WA), with identical font and subheadings, and with tables and figures included at the end of the document. This process was used to eliminate formatting standards, which would make the journal of publication identifiable to the reviewers. Any information that would identify the authors or study location was also removed from the text. Articles were randomly sorted and assigned a unique study identification number for use through the remainder of the review process. Two authors with formal research methodology training independently reviewed each article using a structured data collection instrument that followed the STARD checklist (Table 1). Each item was answered as "yes", "no", or "NA," (not applicable) with opportunity to include comments. If it was thought that a STARD item was not clearly described in the manuscript, "no" was indicated. Disagreements between the two reviewers were adjudicated by a third author, also with formal research methodology training, who was blinded to initial STARD responses. A total STARD score was calculated for each article by summing the total number of STARD criteria satisfied, ranging from 0 to 25.

Data Analysis

All data were manually entered into an electronic spreadsheet (Microsoft Excel; Microsoft, Redmond, WA) and transferred into native SAS format. All analyses were performed with SAS Enterprise Guide Version 5.1 (SAS Institute, Inc., Cary, NC). Descriptive statistics are reported for all variables. Categorical data are reported as percentages with 95% confidence intervals (95% CIs) and continuous data as medians with interquartile ranges (IQRs). Interrater reliability is reported as median raw agreement and kappa statistics with range and IQR. Bivariate comparisons were performed using Fisher's exact test for categorical data and Wilcoxon rank sum test for continuous data. A p-value < 0.05 was considered statistically significant. No a priori sample size calculation was made.

RESULTS

A total of 307 EUS studies were identified, and of these, 45 met criteria for inclusion (Figure 1). Of the 45 articles, only 29 (64%) evaluated the diagnostic accuracy of EUS with calculations of sensitivity and specificity. The remaining articles evaluated EUS as a diagnostic tool, but did not include diagnostic accuracy results.

The median STARD score across all articles was 15 (IQR: 12–17), representing 60% of the all STARD criteria. The median STARD score among articles that reported diagnostic accuracy was significantly higher than those that reported EUS as a diagnostic tool but did not report accuracy (17, IQR: 15–19 vs. 11, IQR: 9–13, respectively; $p < 0.0001$). Of the 45 articles, 32 (50%, 95% CI: 56%−84%) met at least 50% of the STARD criteria, and only 2 (4%, 95% CI: 1%−15%) met greater than 80% of the STARD criteria.

The total number of articles meeting each of the STARD criteria is outlined in Table 2. This table also compares the number of STARD criteria met by the studies that reported diagnostic accuracy information compared to those that did not. The percentage of articles reporting diagnostic accuracy that met each individual STARD criteria is depicted in Figure 2, and of those that did not report diagnostic accuracy is depicted in Figure 3.

Among the articles included in the study, certain STARD criteria were reported the majority of the time. The highest adherence to the STARD criteria, reported 80% or more of the time, was found in STARD criteria 2, 3, 6, 8, 9, 14 and 25 (Table 3). Reporting of some criteria was exceptionally low, reporting 20% or less of the time, in STARD criteria 13, 17, 20 and 24, noting that STARD criteria 20 was met in only a single article (Table 4).

The median observed agreement between the two reviewers across the 25 STARD criteria and 45 reviewed articles was 0.71 (IQR: 0.64–0.80). Kappa was 0.45 (95% CI 0.41–0.50) and the Prevalence Adjusted and Bias Adjusted Kappa was 0.43 (95% CI: 0.34–0.60). Interrater reliability information for each individual criterion is shown in Table 5.

DISCUSSION

An initial study of diagnostic accuracy studies across major medical journals from 1978–1993 demonstrated average methodological quality at best, with publications missing information on how studies were designed, conducted and analyzed[5]. Since the adoption of the STARD criteria, multiple specialties have assessed how well their published studies of diagnostic tests perform with respect to these standards[6–12], reporting overall poor adherence to the STARD guidelines. A more recent study of emergency medicine studies showed variable adherence, ranging from 9–100% of the STARD criteria being met on any given study[13]. There has been just a single study of EUS and STARD, which evaluated the Focused Assessment with Sonography in Trauma (FAST). In this study, a median of 13 methodological criteria (of 27 criteria pooled from STARD and QUADAS) were met, and when methodological standards were missing, investigators significantly overestimated pooled sensitivity of the FAST[14].

Of the studies that dealt with the core EUS applications and the use of ultrasound as a diagnostic tool, we were surprised to find that such a low percentage evaluated diagnostic accuracy and presented sensitivity and specificity information. EUS diagnostic accuracy studies that report findings using the STARD criteria provide the reader the opportunity to confirm and apply the findings of the published study in their specific population. Ideally, all research studies that describe the use of EUS as a diagnostic modality would report diagnostic accuracy data.

Of the articles that we reviewed, a majority reported more than 50% of the STARD criteria, but very few reported more than 80% of the criteria. The median STARD score across all articles was 15, which is 60% of the STARD criteria. Articles that reported diagnostic accuracy information performed slightly better, meeting 17/24, or 68% of the STARD criteria. While this outperforms the reporting of STARD adherence across the literature[7–12,15], it is lower than that of general emergency medicine studies.

In this investigation, we found the reporting of five specific criteria was high. Stating the research aims (STARD 1) and discussing clinical applicability of the test (STARD 25) were included in 100% of the articles. Describing the study population (STARD 3), the technological specifications of the materials and methods (STARD 8) and the process of data collection (STARD 6) were all reported more than 80% of the time. That said, when compared to general emergency medicine diagnostic studies, in which there were 7 categories that were met 100% of the time, and 14 categories that were met >80% of the time, there lies significant room for improvement[13].

The lowest reporting was seen in the reporting of adverse events (STARD 20), reporting the time interval between the index and the reference test (STARD 17), reporting estimates of test reproducibility (STARD 24) and a description of the methods for the calculation of test reproducibility (STARD 13) were all reported less than 20% of the time. These numbers fall well below the lowest percent reporting in these categories in general emergency medicine literature, in which the lowest reported adherence for a single category was 26.1% (Gallo 2017). When the time interval between the index test and the standard is not reported, the reader is unable to determine if there was possibility that the patient's status may have changed or if any treatment was administered between the tests[3]. Reporting test reproducibility and the methods by which this was calculated as variation in this regard can affect diagnostic accuracy. There are often multiple opportunities for variation in reproducibility in interpreting a diagnostic study; as such it is essential that this be reported in order to accurately report diagnostic accuracy[3].

The STARD criteria were initially published to improve the quality of reporting of essential elements of diagnostic accuracy studies to facilitate the critical evaluation of published investigations and enhance the reproducibility of these studies[16]. Certainly, there are likely multiple studies describing diagnostic accuracy in EUS in which the above criteria were met, but not reported[13]. However, when the STARD criteria are not fully reported, readers and colleagues have an incomplete picture of the nature of the study. When readers are unable to evaluate potential adverse events of a new modality, patients are left vulnerable to harm if a study is repeated externally. This review demonstrates that there is significant opportunity for improvement in the reporting of STARD criteria, particularly with respect to reporting of diagnostic accuracy.

This review was carried out shortly after the publication of the initial STARD initiative. As the STARD criteria were initially published in 2003, this investigation was designed to be a baseline study to determine adherence to the STARD criteria immediately following its widespread publication. We hope that in publishing this study we can make EUS investigators more aware of the STARD initiative and enhance adherence, which will result in overall higher quality research. In the future we would like to repeat the review on more recent publications to evaluate for improvement, as well as survey EUS experts on their familiarity and utilization of the STARD criteria. Additionally, the revised STARD criteria were published in 2015 and this study was conducted well ahead of this release. Further evaluation of the most recent EUS literature based on these new criteria is warranted.

LIMITATIONS

This study has a few limitations. The interrater reliability of the initial review showed moderate agreement. We believe this is because we allowed the reviewers to choose “No” or “not applicable” for each article, and the reviewers may have used these differently. While “not applicable” may have applied to some categories in the studies that did not report diagnostic accuracy information, the reviewers may have reported this as “No,” meaning not reported, rather than not applicable. This likely resulted from the significant number of articles reviewed (35%) that did not report diagnostic accuracy information and therefore many of the STARD criteria related to reporting of diagnostic accuracy may have been confusing to classify. Given our initial modest agreement, we did adjudicate all discrepancies with a third reviewer in order to minimize bias of these reviews.

It is possible that some of these studies met some STARD criteria but did not report them all in their paper, or that the intrinsic nature of point of care ultrasound makes it difficult to report certain criteria because of the variability from one user to the next. While we attempted to evaluate all criteria and adjudicate with multiple reviewers, there is the possibility that some criteria could not be evaluated for these reasons.

While we used internet-based search engines and performed a manual search of the table of contents of each major journal, we may have inadvertently omitted diagnostic accuracy articles from the study. To help mitigate this possibility, we used a stepwise approach to identification of articles for inclusion and also conducted a PubMed search to help identify any additional articles meeting study criteria.

Finally, articles included in our study ranged from 2005 – 2010 and methodological rigor of EUS literature may have improved since 2010.

CONCLUSIONS

Significant opportunities exist to improve methodological reporting of EUS research. Increased adherence to the STARD criteria among diagnostic EUS studies is necessary to improve reporting and ability to compare outcomes.

REFERENCES

1. Lijmer JG. Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests. JAMA 1999;282(11):1061. [PubMed: 10493205]
2. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33(1):85–91. [PubMed: 9867892]
3. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49(1):7–18. [PubMed: 12507954]
4. Policy Statement: Emergency Ultrasound Guidelines. Annals of Emergency Medicine 2009;53(4):550–70. [PubMed: 19303521]
5. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274(8):645–51. [PubMed: 7637146]
6. Coppus S, Vanderveen F, Bossuyt P, Mol B. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. Fertility and Sterility 2006;86(5):1321–9. [PubMed: 16978620]
7. Michelessi M, Lucenteforte E, Miele A, et al. Diagnostic accuracy research in glaucoma is still incompletely reported: An application of Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. PLOS ONE 2017;12(12):e0189716. [PubMed: 29240827]
8. Roposch A, Moreau NM, Uleryk E, Doria AS. Developmental Dysplasia of the Hip: Quality of Reporting of Diagnostic Accuracy for US. Radiology 2006;241(3):854–60. [PubMed: 17053199]
9. Siddiqui MAR. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. British Journal of Ophthalmology 2005;89(3):261–5. [PubMed: 15722299]
10. Smidt N, Rutjes AWS, van der Windt DAWM, et al. Quality of Reporting of Diagnostic Accuracy Studies. Radiology 2005;235(2):347–53. [PubMed: 15770041]
11. Selman TJ, Khan KS, Mann CH. An evidence-based approach to test accuracy studies in gynecologic oncology: the ‘STARD’ checklist. Gynecologic Oncology 2005;96(3):575–8. [PubMed: 15721396]
12. Selman TJ, Morris RK, Zamora J, Khan KS. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. BMC Women’s Health [Internet] 2011 [cited 2019 Apr 16];11(1). Available from: http://bmcwomenshealth.biomedcentral.com/articles/10.1186/1472-6874-11-8
13. Gallo L, Hua N, Mercuri M, Silveira A, Worster A, Best Evidence in Emergency Medicine (BEEM; beem.ca). Adherence to Standards for Reporting Diagnostic Accuracy in Emergency Medicine Research. Academic Emergency Medicine 2017;24(8):914–9. [PubMed: 28621810]
14. Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between Compliance with Methodological Standards of Diagnostic Research and Reported Test Accuracy: Meta-Analysis of Focused Assessment of US for Trauma. Radiology 2005;236(1):102–11. [PubMed: 15983072]
15. Korevaar DA, Wang J, van Enst WA, et al. Reporting Diagnostic Accuracy Studies: Some Improvements after 10 Years of STARD. Radiology 2015;274(3):781–9. [PubMed: 25350641]
16. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;h5527. [PubMed: 26511519]

Article Summary:

1. Why is this topic important?

   a. Emergency point-of-care ultrasound is widely used as a diagnostic tool in emergency departments to guide clinical decision making. Given this, rigorous study surrounding its accuracy is essential.

2. What does this study attempt to show?

   a. This study attempts to demonstrate how publications describing EUS diagnostic studies adhered to the STARD criteria after their initial widespread distribution.

3. What are the key findings?

   a. The median STARD score was 15 (IQR: 12 – 17), which is 60% of the 25 total STARD criteria. The median STARD score among articles that reported diagnostic accuracy was significantly higher than those that did not report diagnostic accuracy [17 IQR: 15 – 19) vs. 11 (IQR: 9 – 13), respectively; $p < 0.0001$]. Seventy one percent of articles met at least 50% of the STARD criteria (56%−84%) and 4% met greater than 80% of the STARD criteria.

4. How is patient care impacted?

   a. We aim to advance the quality of emergency ultrasound diagnostic studies, which in turn will advance the science and clinical practice of emergency medicine, and therefore individual patient care.

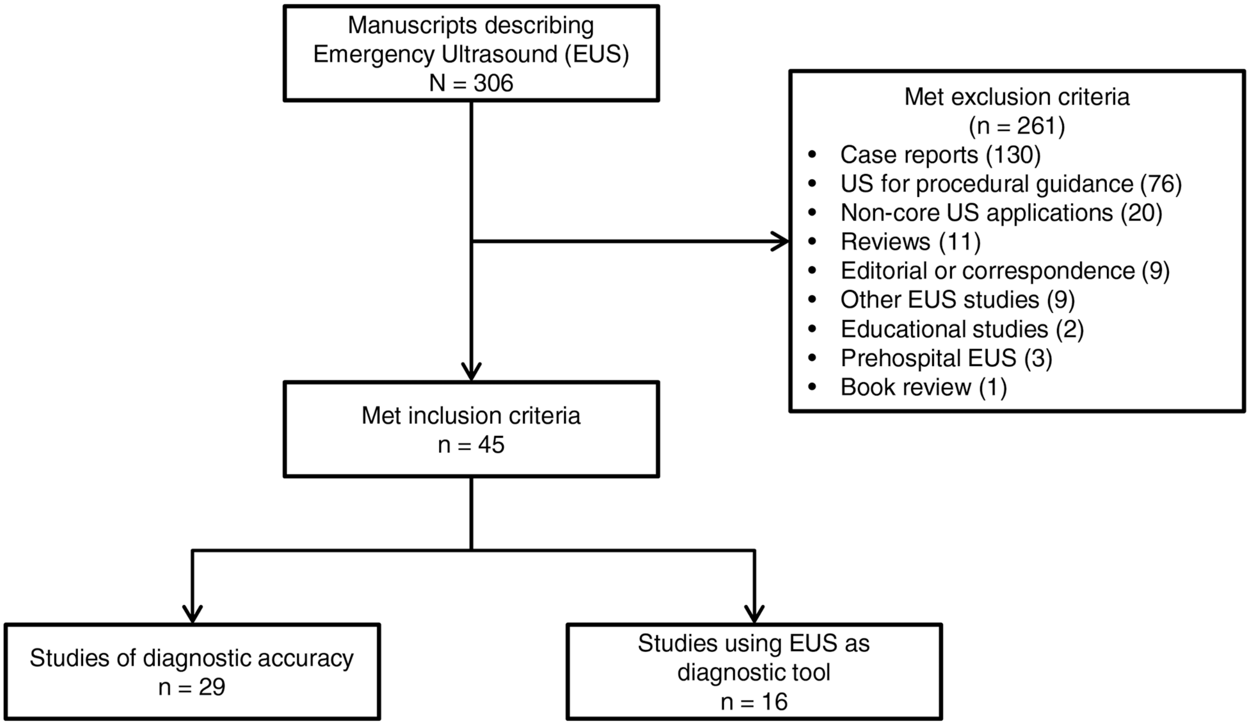
Manuscripts describing Emergency Ultrasound (EUS)
N = 306
Met exclusion criteria
(n = 261)
• Case reports (130)
• US for procedural guidance (76)
• Non-core US applications (20)
• Reviews (11)
• Editorial or correspondence (9)
• Other EUS studies (9)
• Educational studies (2)
• Prehospital EUS (3)
• Book review (1)
Met inclusion criteria
n = 45
Studies of diagnostic accuracy
n = 29
Studies using EUS as diagnostic tool
n = 16

Figure 1.

Flow diagram of selected articles.

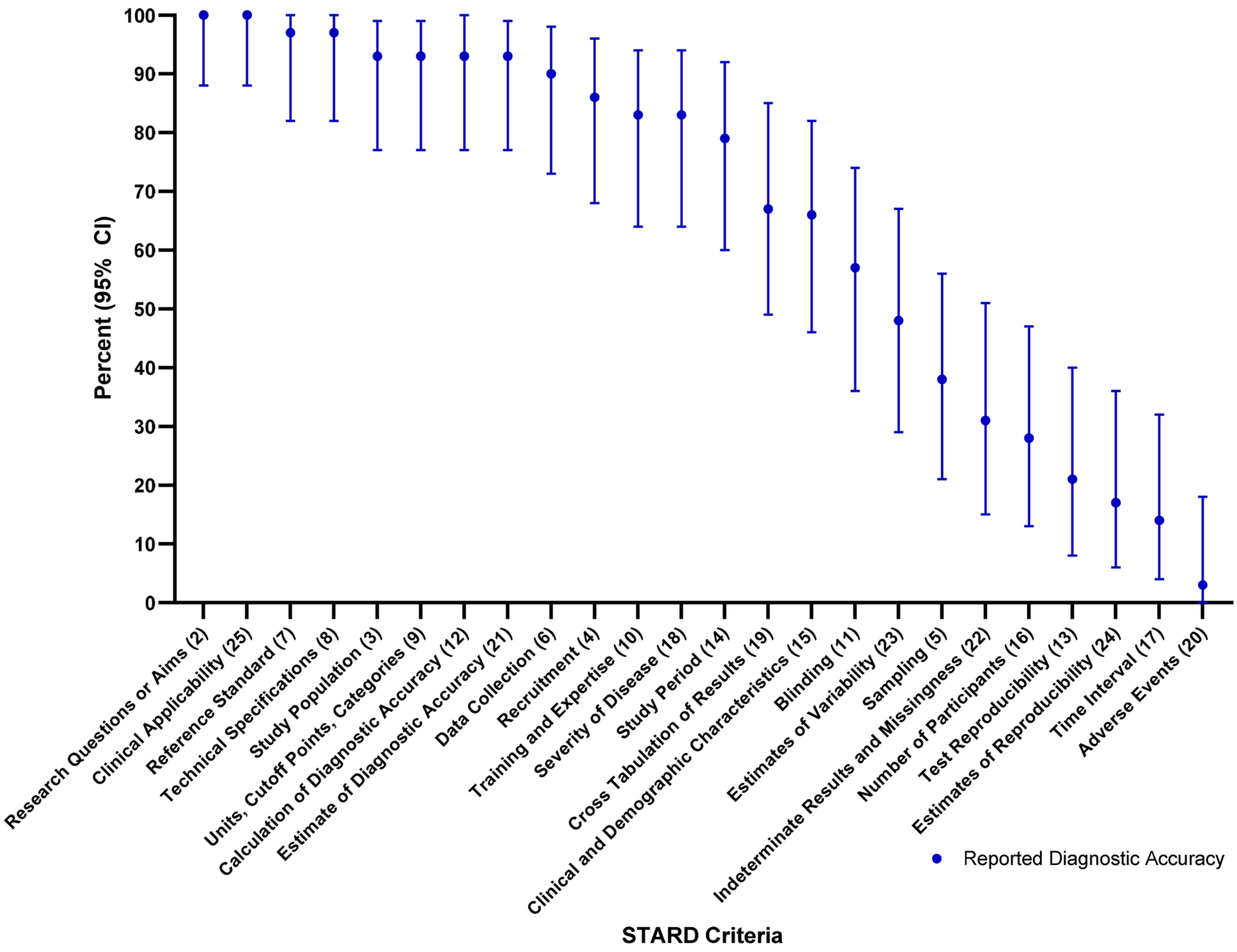

Figure 2.

STARD criteria for articles that reported diagnostic accuracy.

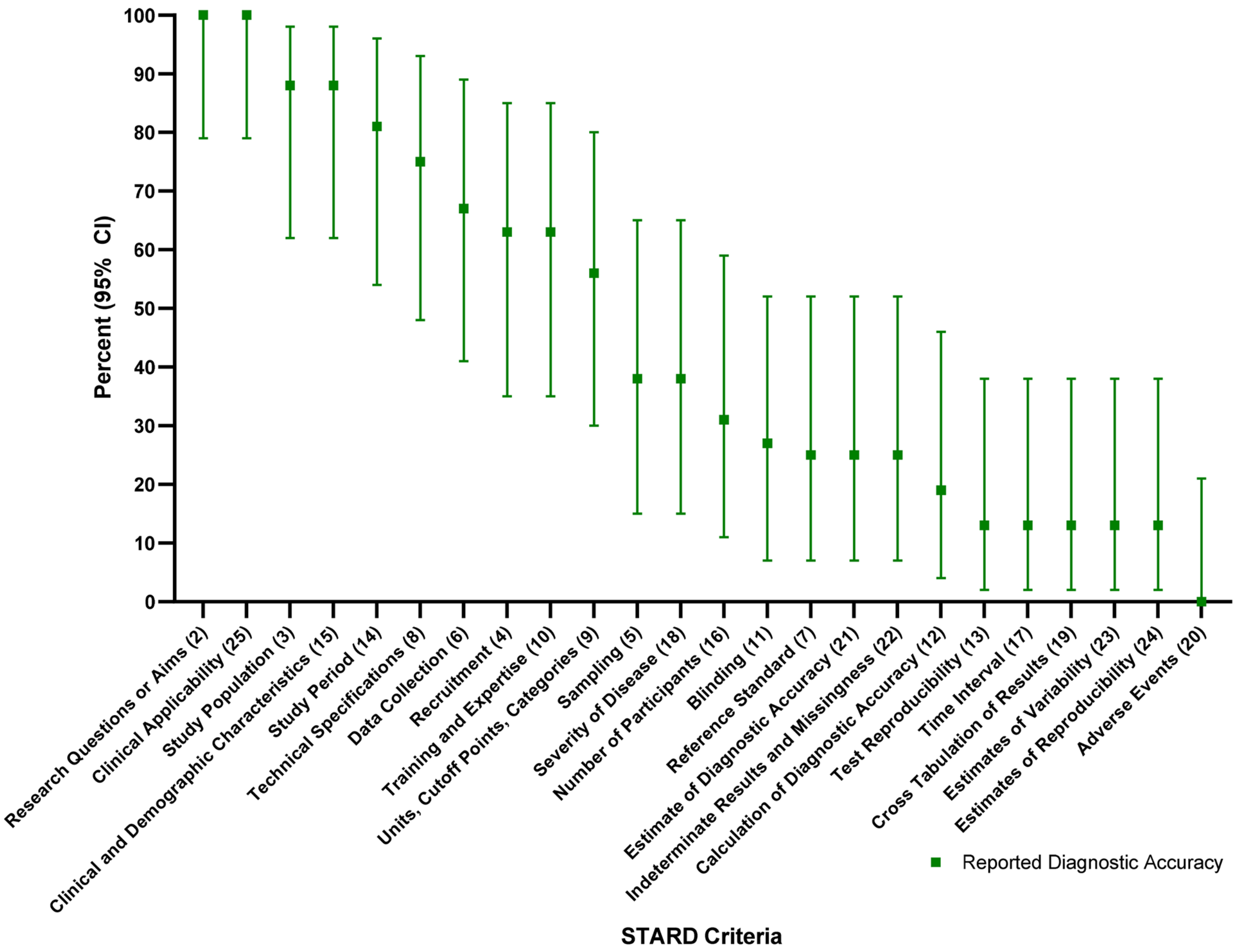

Figure 3.

STARD Criteria for articles that did not report diagnostic accuracy.

Table 1.

STARD checklist for the reporting of studies of diagnostic accuracy (first official version, January 2003)

Section and Topic	**Item #**	
TITLE/ASTRACT/ KEWORDS	1	Is the article a study of diagnostic accuracy?
INTRODUCTION	2	State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups
METHODS		
Participants	3	Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected
	4	Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Describe participant sampling: was this a consecutive series of participants defined by selection criteria in the previous 2 items? If not, specify how participants were further selected
	6	Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?
Test Methods	7	Describe the reference standard and its rationale
	8	Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both
	9	Describe definition of and rationale for the units, cutoff points, or categories of the results of the index tests and the reference standard
	10	Describe the number, training, and expertise of the persons executing and reading the index tests and the reference standard
	11	Were the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to the readers.
Statistical Methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (eg 95% confidence intervals)
	13	Describe methods for calculating test reproducibility, if done
RESULTS		
Participants	14	Report when study was done, including beginning and ending dates of recruitment
	15	Report clinical and demographic characteristics (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment center)
	16	Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test Results	17	Report time interval from index tests to reference standard, and any treatment administered between
	18	Report distribution of severity of disease (define criteria) in those with the target condition and other diagnoses in participants without the target condition
	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, report the distribution of the test results by the results of the reference standard
	20	Report any adverse events from performing the index test or the reference standard
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)
	22	Report how indeterminate results, missing responses, and outliers of index tests were handled
	23	Report estimates of variability of diagnostic accuracy between readers, centers, or subgroups of participants, if done
	24	Report estimates of test reproducibility, if done

Section and Topic	**Item #**	
DISCUSSION	25	Discuss the clinical applicability of the study findings

Table 2.

Articles reporting diagnostic accuracy compared to those that did not report diagnostic accuracy.

STARD CRITERIA	Reported Accuracy n = 29 (64, 49 – 78)	Did Not Report Accuracy n = 16 (36, 22 – 51)	All Articles N = 45
	n (%, 95% CI)	n (%, 95% CI)	n (%, 95% CI)
2. State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	29 (100, 88 – 100)	16 (100, 79 – 100)	45 (100, 92 – 100)
3. Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected	27 (93, 77 – 99)	14 (88, 62 – 98)	41 (91, 79 – 98)
4. Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	25 (86, 68 – 96)	10 (63, 35 – 85)	35 (78, 63 – 89)
5. Describe participant sampling: was this a consecutive series of participants defined by selection criteria in the previous 2 items? If not, specify how participants were further selected	11 (38, 21 – 56)	6 (38, 15 – 65)	17 (38, 24 – 54)
6. Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?	26 (90, 73 – 98)	11 (67, 41 – 89)	37 (82, 68 – 92)
7. Describe the reference standard and its rationale	28 (97, 82 – 100)	4 (25, 7 – 52)	32 (71, 56 – 84)
8. Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both	28 (97, 82 – 100)	12 (75, 48 – 93)	40 (89, 76 – 96)
9. Describe definition of and rationale for the units, cutoff points, or categories of the results of the index tests and the reference standard	27 (93, 77 – 99)	9 (56, 30 – 80)	36 (80, 65 – 90)
10. Describe the number, training, and expertise of the persons executing and reading the index tests and the reference standard	24 (83, 64 – 94)	10 (63, 35 – 85)	34 (76, 60 – 87)
11. Were the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to the readers.	16 (57, 36 – 74)	4 (27, 7 – 52)	20 (44, 30 – 60)
12. Describe methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (eg 95% confidence intervals)	27 (93, 77 – 100)	3 (19, 4 – 46)	30 (67, 51 – 80)
13. Describe methods for calculating test reproducibility, if done	6 (21, 8 – 40)	2 (13, 2 – 38)	8 (18, 8 – 32)
14. Report when study was done, including beginning and ending dates of recruitment	23 (79, 60 – 92)	13 (81, 54 – 96)	14 (80, 65 – 90)
15. Report clinical and demographic characteristics (eg age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centre)	19 (66, 46 – 82)	14 (88, 62 – 98)	33 (73, 58 – 85)
16. Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)	8 (28, 13 – 47)	5 (31, 11 – 59)	13 (29, 16 – 44)
17. Report time interval from index tests to reference standard, and any treatment administered between	4 (14, 4 – 32)	2 (13, 2 – 38)	6 (13, 5 – 27)
18. Report distribution of severity of disease (define criteria) in those with the target condition and other diagnoses in participants without the target condition	24 (83, 64 – 94)	6 (38, 15 – 65)	30 (67, 50 – 80)
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, report the distribution of the test results by the results of the reference standard	20 (67, 49 – 85)	2 (13, 2 – 38)	22 (49, 34 – 64)

STARD CRITERIA	Reported Accuracy n = 29 (64, 49 – 78)	Did Not Report Accuracy n = 16 (36, 22 – 51)	All Articles N = 45
	n (%, 95% CI)	n (%, 95% CI)	n (%, 95% CI)
20. Report any adverse events from performing the index test or the reference standard	1 (3, 0 – 18)	0 (0, 0 – 21)	1 (2, 0 – 12)
21. Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)	27 (93, 77 – 99)	4 (25, 7 – 52)	31 (67, 53 – 82)
22. Report how indeterminate results, missing responses, and outliers of index tests were handled	9 (31, 15 – 51)	4 (25, 7 – 52)	13 (29, 16 – 44)
23. Report estimates of variability of diagnostic accuracy between readers, centers, or subgroups of participants, if done	14 (48, 29 – 67)	2 (13, 2 – 38)	16 (36, 22 – 51)
24. Report estimates of test reproducibility, if done	5 (17, 6 – 36)	2 (13, 2 – 38)	7 (16, 6 – 29)
25. Discuss the clinical applicability of the study findings	29 (100, 88 – 100)	16 (100, 79 – 100)	45 (100, 92 – 100)

Table 3.

STARD Criteria with Highest Adherence (80% or more)

STARD CRITERIA	Reported Accuracy n = 29 (64, 49 – 78)	Did Not Report Accuracy n = 16 (36, 22 – 51)	All Articles N = 45
	n (%, 95% CI)	n (%, 95% CI)	n (%, 95% CI)
2. State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	29 (100, 88 – 100)	16 (100, 79 – 100)	45 (100, 92 – 100)
3. Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected	27 (93, 77 – 99)	14 (88, 62 – 98)	41 (91, 79 – 98)
6. Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?	26 (90, 73 – 98)	11 (67, 41 – 89)	37 (82, 68 – 92)
8. Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both	28 (97, 82 – 100)	12 (75, 48 – 93)	40 (89, 76 – 96)
9. Describe definition of and rationale for the units, cutoff points, or categories of the results of the index tests and the reference standard	27 (93, 77 – 99)	9 (56, 30 – 80)	36 (80, 65 – 90)
14. Report when study was done, including beginning and ending dates of recruitment	23 (79, 60 – 92)	13 (81, 54 – 96)	36 (80, 65 – 90)
25. Discuss the clinical applicability of the study findings	29 (100, 88 – 100)	16 (100, 79 – 100)	45 (100, 92 – 100)

Table 4.

STARD Criteria with Lowest Adherence (20% or less)

STARD CRITERIA	Reported Accuracy n = 29 (64, 49 – 78)	Did Not Report Accuracy n = 16 (36, 22 – 51)	All Articles N = 45
	n (%, 95% CI)	n (%, 95% CI)	n (%, 95% CI)
13. Describe methods for calculating test reproducibility, if done	6 (21, 8 – 40)	2 (13, 2 – 38)	8 (18, 8 – 32)
17. Report time interval from index tests to reference standard, and any treatment administered between	4 (14, 4 – 32)	2 (13, 2 – 38)	6 (13, 5 – 27)
20. Report any adverse events from performing the index test or the reference standard	1 (3, 0 – 18)	0 (0, 0 – 21)	1 (2, 0 – 12)
24. Report estimates of test reproducibility, if done	5 (17, 6 – 36)	2 (13, 2 – 38)	7 (16, 6 – 29)

Table 5:

Inter-Rater Reliability

	Observed Agreement	**Cohen’s Kappa (95% CI)**	**Prevalence Adjusted Bias Adjusted Kappa**
Inter-Rater Reliability	0.71	0.45 (0.41–0.50)	0.43
IRR STARD Questions 1			
1	0.80	0.59 (0.36–0.82)	0.60
2	0.96	0.00 (0–0)	0.91
3	0.76	0.25 (−0.03 to 0.53)	0.51
4	0.71	0.32 (0.09–0.55)	0.42
5	0.29	0.05 (−0.01 to 0.10)	−0.42
6	0.71	0.06 (−0.15 to 0.27)	0.42
7	0.53	0.23 (0.08–0.38)	0.07
8	0.49	0.10 (−0.01 to 0.21)	−0.02
9	0.51	0.13 (0.0–0.26)	0.02
10	0.56	0.21 (0.20–0.39)	0.11
11	0.71	0.41 <0.15–0.67)	0.42
12	0.69	0.39 (0.14–0.64)	0.38
13	0.82	0.39 (0.05–0.74)	0.64
14	0.89	0.71 (0.47–0.95)	0.78
15	0.67	0.36 (0.16–0.56)	0.33
16	0.64	0.28 (0.06–0.50)	0.29
17	0.78	0.26 (−0.06 to 0.58)	0.56
18	0.71	0.43 (0.20–0.67)	0.42
19	0.69	0.36 (0.09–0.64)	0.38
20	0.93	−0.03 (−0.07 to 0.01)	0.87
21	0.80	0.61 (0.39–0.83)	0.60
22	0.62	0.23 (0.05–0.42)	0.24
23	0.76	0.35 (0.08–0.63)	0.51
24	0.80	0.30 (−0.04 to 0.63)	0.60
25	1.00	1.00 (1–1)	1.00