Reporting standards for studies of diagnostic test accuracy in dementia

The STARDdem Initiative

Anna H. Noel-Storr, MSc
Jenny M. McCleery, MB, BS
Edo Richard, MD
Craig W. Ritchie, MD
Leon Flicker, MD
Sarah J. Cullum, MBChB, MRCPsych, PhD
Daniel Davis, MD
Terence J. Quinn, MD
Chris Hyde, MBBS
Anne W.S. Rutjes, PhD
Nadja Smailagic, MD
Sue Marcus, MSc
Sandra Black, MD
Kaj Blennow, MD
Carol Brayne, MD
Mario Fiorivanti, MD
Julene K. Johnson, PhD
Sascha Köpke, PhD
Lon S. Schneider, MD
Andrew Simmons, PhD
Niklas Mattsson, MD
Henrik Zetterberg, MD
Patrick M.M. Bossuyt, PhD
Gordon Wilcock, DM
Rupert McShane, MD

Correspondence to Dr. McShane: rupert.mcshane@oxfordhealth.nhs.uk

ABSTRACT

Objective: To provide guidance on standards for reporting studies of diagnostic test accuracy for dementia disorders.

Methods: An international consensus process on reporting standards in dementia and cognitive impairment (STARDdem) was established, focusing on studies presenting data from which sensitivity and specificity were reported or could be derived. A working group led the initiative through 4 rounds of consensus work, using a modified Delphi process and culminating in a face-to-face consensus meeting in October 2012. The aim of this process was to agree on how best to supplement the generic standards of the STARD statement to enhance their utility and encourage their use in dementia research.

Results: More than 200 comments were received during the wider consultation rounds. The areas at most risk of inadequate reporting were identified and a set of dementia-specific recommendations to supplement the STARD guidance were developed, including better reporting of patient selection, the reference standard used, avoidance of circularity, and reporting of test-retest reliability.

Conclusion: STARDdem is an implementation of the STARD statement in which the original checklist is elaborated and supplemented with guidance pertinent to studies of cognitive disorders. Its adoption is expected to increase transparency, enable more effective evaluation of diagnostic tests in Alzheimer disease and dementia, contribute to greater adherence to methodologic standards, and advance the development of Alzheimer biomarkers. *Neurology*® 2014;83:364–373

GLOSSARY

AD = Alzheimer disease; **CONSORT** = Consolidated Standards of Reporting Trials; **DTA** = diagnostic test accuracy; **MCI** = mild cognitive impairment; **STARD** = Standards for Reporting of Diagnostic Accuracy.

Over the past decade, there has been an impressive increase in the number of reports published on Alzheimer disease (AD) and dementia biomarkers, describing both proof-of-concept and diagnostic test accuracy (DTA) studies. New diagnostic criteria proposed in the United States and in Europe place greater emphasis on the use of biomarkers and imaging techniques in the diagnosis of AD (in both symptomatic and asymptomatic subjects).[1–4] An amyloid PET ligand has been licensed on the basis of its utility in excluding a diagnosis of AD.[5,6] "Appropriate Use" criteria have since been proposed that suggest routine use of florbetapir PET scanning in mild cognitive impairment (MCI), but the lack of evidence for enhancing diagnostic certainty or for predicting progression has been acknowledged.[7] The European Medicines Agency supports the use of CSF biomarkers (β-amyloid 42 and tau) to enrich clinical populations with prodromal AD.[8] Numerous other potential biomarkers are in development.

Diagnostic tests for diseases that may cause cognitive problems are not restricted to biochemical and neuroimaging biomarkers. There are a variety of clinical assessment scales, both for "screening" and "diagnosis," such as the Alzheimer's Disease Assessment Scale,[9] Montreal Cognitive Assessment,[10] and the Addenbrooke's Cognitive Examination–Revised.[11] At present, there is little guidance on the optimal assessment scale for a particular purpose or setting; this has resulted in considerable variation in approaches to cognitive testing. As many countries move toward large-scale cognitive screening of older adults,[12] there is considerable need for studies of DTA.[13,14]

Supplemental data at Neurology.org

Authors' affiliations are listed at the end of the article.

Go to Neurology.org for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

In dementia, diagnostic studies can be divided into proof-of-concept studies (whether a test result is different in healthy controls vs dementia patients) and studies investigating the clinical applicability of a new diagnostic test. For the latter, there are 3 main questions:

1. Are those with certain test results more likely to have a particular form of dementia, for example, Lewy body dementia, than persons with other test results (differential diagnosis)?
2. Are those with certain test results more likely to progress to, for example, AD dementia, than persons with other test results (delayed determination or prediction)?
3. Does the test provide incremental benefit in the diagnostic workup, considering ease of administration of the test, costs, and burden for the patient?

In this context, the quality of reporting of DTA studies is particularly important. Poor or inconsistent reporting can hamper effective evaluation of study methods, assessment of potential for bias, and interpretation of the results.[15–18] It also limits the ability to synthesize data across studies, precluding methodologically sound meta-analyses. Guidelines for reporting standards in other contexts,[19] such as the Consolidated Standards of Reporting Trials (CONSORT) statement, are effective in raising reporting standards and, indeed, can also drive improvements in standards of trial design.[20–22]

In 2003, Bossuyt et al.[23] published the STARD statement: Standards for Reporting of Diagnostic Accuracy studies, aiming to "improve the accuracy and completeness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study (internal validity) and to evaluate its generalizability (external validity)."[24] To date, more than 200 scientific journals have become STARD "adopters." Although the impact of STARD may not yet be comparable to CONSORT, STARD has raised standards of reporting in diagnostic accuracy studies.[25]

Despite this, a recent systematic review found that, within the field of dementia, the majority of reports of diagnostic biomarker studies were missing important information, particularly for blinding of results of either the biomarker or reference standard, handling of missing data, sample selection methods, and test reproducibility.[26] Although DTA studies in dementia are similar to those in any field, some features specific to dementia research are not fully addressed in the STARD criteria. Therefore, we aimed to identify aspects of reporting that are particularly important in the context of AD and dementia, produce dementia-specific supplementary guidance to STARD, and thereby enhance use (and utility) of STARD for dementia studies.

METHODS The STARDdem Initiative sought to establish an international consensus on reporting standards for DTA studies in dementia, highlighting the most important issues and identifying any areas in which additional reporting recommendations might enhance the usefulness of the STARD guideline. The method used was derived from guidance on effective implementation of reporting standards published by the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network,[27] which proposes a step-wise development approach and deems a consensus process to be "a crucial characteristic" of developing reporting guidelines.[22] We adapted the guidance leading to a development process comprising 3 broad phases: (1) evaluation, (2) drafting with widespread discussion and feedback using a modified Delphi technique,[28] and (3) delivery. This report describes the first 2 phases and itself constitutes part of the delivery phase.

Phase 1: Evaluation. We conducted a comprehensive literature review on biomarker DTA studies in dementia, focusing on studies that included patients with cognitive impairment (but no dementia) at baseline and used progression to dementia of the AD type as a reference standard (for methods, see Noel-Storr et al.[26]). Applying STARD, we assessed the quality of reporting in all identified DTA studies by calculating the percentage of studies complying with each of the STARD items.[23]

Phase 2: Drafting and discussion. An international and multidisciplinary working group of dementia experts and methodologists was established, organized by the Cochrane Dementia and Cognitive Improvement Group.[29] The objectives of this group were to (1) define the scope of STARD for dementia, (2) assess the applicability of existing reporting guidelines (STARD) to dementia, (3) draft dementia-specific supplementary recommendations,

Figure Phases and rounds of the STARDdem Consensus Initiative

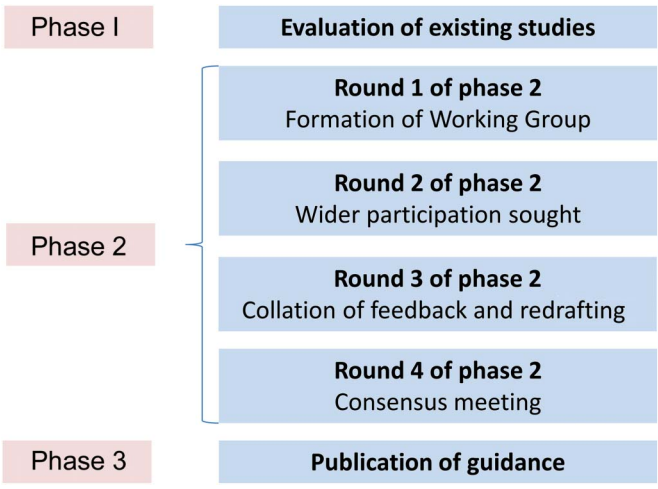

Table 1 Sources of bias

Bias	Explanation	Example
Test performance bias		
Context bias	Readers are more likely to interpret results from (subjective) tests as abnormal in settings with higher prevalence of the target condition	In an early-onset dementia clinic with a high prevalence of FTD, apathy may be more likely to be seen as supportive of this diagnosis than in an older population
Clinical review bias	Interpretations are influenced by providing additional clinical information to interpreters	When clinical information about cognitive function may influence a radiologist's assessment of hippocampal atrophy
Diagnostic review bias	Knowing the result of the index test while interpreting the reference standard. Leads to inflated diagnostic accuracy	When knowledge of a CSF Aβ and/or tau test result may influence a clinician making a diagnosis of AD
Verification bias	If the result of the index test influences the decision to order the reference standard test	Common in dementia research. Would occur if PiB-PET is the index test, AD at autopsy is the reference standard, and the decision to conduct an autopsy was (partly) based on the PET result
Incorporation bias	If the index test forms part of the reference standard, this leads to circularity and inflated diagnostic accuracy	Very common in dementia research, for example, when the index test is a test of episodic memory and the reference standard is a clinical diagnosis of AD
Test review bias	Knowledge of the result of the reference standard while interpreting the index test. Leads to inflated diagnostic accuracy	More common in cross-sectional studies. For example, knowledge of a patient's diagnosis may influence a radiologist's interpretation of an amyloid PET scan
Patient-based bias		
Limited challenge bias	Patients with a condition known to adversely affect the way the test works are excluded. Leads to inflated diagnostic accuracy	Common. Occurs when patients likely to be more challenging to diagnose are excluded, for example, applying a potential index test for DLB to a sample from which "possible DLB" subjects have been excluded. (Could also be an example of spectrum bias—see below)
Selection bias	If participants are excluded based on specific characteristics	Common. Occurs, for example, when patients with cerebrovascular lesions on MRI are excluded from a study investigating a biomarker for AD although the population to whom the test may be applied in daily practice will include many patients with some cerebrovascular lesions
Spectrum bias	Demographic and clinical features, including disease severity of the study population, influence diagnostic accuracy	Common. For example, studying a relatively young group of patients (e.g., <70 y) who are likely to differ in many ways from the older population who constitute the majority of patients with dementia

Abbreviations: Aβ = β-amyloid; AD = Alzheimer disease; DLB = dementia with Lewy bodies; FTD = frontotemporal dementia; PiB = Pittsburgh compound B.

and (4) seek feedback and consensus from the dementia research community.

In round 1 of phase 2 (see the figure), the working group held a series of 4 meetings (one face-to-face, 3 teleconferences). Before each, members independently assessed 3 papers using the STARD tool, rating whether each STARD criterion was met and with the option to record free text comments. The 12 papers were randomly selected from studies identified by searches for DTA systematic reviews in progress by the Cochrane Dementia and Cognitive Improvement Group.[30] Omissions, agreements, and disagreements were identified and discussed at these meetings. This stage of the process highlighted areas in which there was lack of clarity about what constituted "clear" or "poor" reporting within a dementia context, and hence identifying a number of focus areas for STARD-dem. Supplementary guidance in these areas was drafted, and examples of adequate/clear reporting were identified from dementia diagnostic research for each relevant item. Round 1 also helped to highlight the different types of bias that can arise in dementia studies (see table 1). Bias, in this context, is defined as a systematic error, often unintentional and sometimes unavoidable, in an observed measurement from the true value. If bias existed, a study would consistently over- or underestimate the true accuracy parameters (such as test sensitivity or specificity) were the study to be replicated and repeated.

In round 2 (figure), draft additional guidance and examples, together with the generic STARD guidance, were uploaded to the STARDdem Web site.[31] More than 350 individuals were invited to comment via the Web site. These individuals had been identified as the following: corresponding authors of diagnostic studies using imaging, biochemical biomarkers, or neuropsychiatric

Table 2 STARD items poorly reported or referenced

Item no.	Topic	Partial or not reported, %
5	Participant sampling	46
10	Training and expertise	
	Index test	73
	Reference standard	77
11	Blinding	
	Index test	77
	Reference standard	36
13	Methods for calculating test reproducibility	
	Index test	76
	Reference standard (operationalized)	96
14	Study dates	72
16	Reasons for subjects meeting inclusion who did not undergo index test or reference standard	60
20	Presence or absence of adverse events	97
22	Handling of missing or indeterminate data	82
23	Variability of diagnostic accuracy between subgroups: participants, readers, or centers	64
24	Estimates of test reproducibility	
	Index test	82
	Reference standard	96

tests; presenting authors of relevant oral presentations and posters identified from the abstract books of the Alzheimer's Association International Conference 2012 and Clinical Trials in Alzheimer's Disease conference 2012; editors of journals who publish significant numbers of diagnostic studies in dementia; and DTA methodologists. The site was open access with an encouraged branching dissemination strategy whereby participants shared the site address with other interested parties. Participants could post general feedback and/or comments on specific items, anonymously if they wished. This period of feedback was open for 2 months. The comments obtained were all discussed in the working group during a further 3 teleconferences, and revised dementia-specific additions to STARD were drafted (round 3, figure).

Round 4 (figure) consisted of a half-day consensus meeting, held in October 2012. The meeting had 2 main aims: (1) to reach consensus on the reporting of items of key significance in dementia DTA studies and on the choice of examples to illustrate good reporting, and (2) to decide on the best method for disseminating the outcome of the process and ensuring its adoption by the research community. The meeting ran as both a face-to-face meeting and Web conference to try to maximize attendance and participation. Participants comprised 40 individuals, with key groups represented by researchers/authors in this field and journal editors. After this meeting, the working group then produced a final version of the STARDdem supplementary recommendations.

RESULTS See appendix (table e-2) on the *Neurology*® Web site at Neurology.org for key definitions pertinent to studies of DTA.

Phase 1. The results of the STARD assessment for studies included in our systematic review have been reported in detail elsewhere.[26] In brief, this review found that of the 142 studies identified, there was marked variation in the quality of reporting between several STARD items. Items particularly poorly reported or referenced are listed in table 2.

Phase 2. STARD is applicable to studies in which the results of one or more (index) tests are compared with the results of a reference standard applied to the same subjects, allowing production of a 2×2 table from which estimates of test accuracy—usually sensitivity and specificity—may be obtained. Binary diagnostic categories retain core utility as the basis for prognosis, treatment, management, and legal decision-making: the fact that the pathology and symptoms of dementias occur on a spectrum does not negate the need for the clear delineation of thresholds in index tests and for categorical reference standards. Correlations between continuous variables (e.g., biomarker level and cognitive decline) are of value for establishing etiology and point to potential as a diagnostic test, but do not guide clinicians or reimbursers about when the benefits of starting treatment outweigh the risks and costs.

Although most diagnostic test studies are cross-sectional by design, in dementia studies, the reference standard of "progression from MCI to dementia" is frequently used. These "delayed verification" studies (entailing some additional complexities of design) were included if 2×2 data were presented or could be derived. Studies were included regardless of the phase of development of the test.

The initial draft of reporting items, produced by the working group for consultation, may be viewed on the STARDdem Web site.[31] During the open comment period of 2 months, more than 200 comments were posted by clinicians, statisticians, methodologists, neuropsychologists, molecular biologists, clinical chemists, and radiologists. Based on the comments of the working group, the most important items to address were (1) the description of the population under study, (2) reporting of the operationalization and application of the reference standard, (3) identification of potential incorporation bias or "circularity" when the index test forms a part of the reference standard (e.g., a neuropsychological test that also contributes to the diagnosis of dementia), and (4) reporting of test-retest reliability.

The second draft, which was circulated to consensus meeting participants, may also be viewed on the STARDdem Web site.[31] After the discussion at the consensus meeting, a final version was produced in

Table 3 STARDdem checklist for the reporting of diagnostic accuracy studies in dementia

Section, topic, and item no.	STARD checklist item	Points of particular relevance to dementia
Title/abstract/ keywords		
1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity")	Studies reporting a sensitivity/specificity or 2 × 2 data derivable fall within the scope of STARDdem and should be indexed accordingly
Introduction		
2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	Some studies describing aims related to "prognosis" or "prediction" may also fall within the remit of STARDdem. Report test purpose: "stand-alone" test or as an addition to other tests or clinical criteria
Methods		
Participants		
3	The study population: the inclusion and exclusion criteria, setting and locations where data were collected. See also item 4 on recruitment and item 5 on sampling	Key inclusion criteria: (1) demographic, especially age; (2) cognition- or disease-related criteria. Accurate description of the target sample is required including reporting criteria used to define the study population. Report referral pathways, precise locations of patient recruitment, where index test and reference standard were performed. For secondary/ tertiary settings, helpful to report the medical subspecialty or hospital department (e.g., psychiatry, neurology). Diagnostic accuracy studies in dementia are often nested within larger cohort studies. If this is the case, then the targeted population for the cohort study and the method of selection into the cohort should be described and/or the parent study cited
4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard? See also item 5 on sampling and item 16 on participant loss at each stage of the study	Report whether those in intermediate categories (e.g., possible AD or possible DLB) were excluded
5	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. See also item 4 on recruitment and item 16 on participant loss	Planned analyses showing how characteristics of the subgroup entering the study differ from the eligible population are strongly recommended (i.e., if a convenience sample has been used because of the invasive nature of the test or tests)
6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	Authors should report the timing of the analysis plan regarding data collection: Was the analysis plan set out in a protocol before index and reference standards were performed? If not, when was the analysis plan created?
Test methods		
7	The reference standard and its rationale	For neuropathologic and clinical reference standards, the diagnostic criteria used should be specified. Where relevant, reference should be made to studies validating the criteria. Report whether standard consensus clinical criteria incorporate the index test (incorporation bias rendering blinding of index test impossible)
8	Technical specifications of materials and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard. See also item 10 concerning the person(s) executing the tests	Use of scales: specify details of administration, which version. Clinical diagnostic criteria: what information was available to inform the diagnoses; how the criteria were applied (e.g., by individual clinicians, by consensus conference, by semiautomated algorithm). Imaging and laboratory tests: specify materials and instruments, including sample handling and concordance with any harmonization criteria. In new assays, describe all steps in detail. Any particular preparation of participants should be described
9	Definition of and rationale for the units, cutoffs, and/or categories of the results of the index tests and the reference standard	Justify any cutoffs used, because these may vary with clinical context

Continued

Table 3 Continued

Section, topic, and item no.	STARD checklist item	Points of particular relevance to dementia
10	The number, training, and expertise of the persons executing and reading the index tests and the reference standard. See also item 8	Especially where subjective judgments are involved, e.g., the interpretation of neuroimaging results. Report inter- and intrarater agreement. Reference or describe the content of training materials used. Reference or describe details of lab certification and harmonized biomarker assays
11	Whether or not the readers of the index tests and reference standard were blinded (masked) to the results of the other test and describe any other clinical information available to the readers. See also item 7	Also, the index test may form a part of the reference standard. This is often referred to as incorporation bias and renders blinding of the index test impossible
Statistical methods		
12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals)	
13	Methods for calculating test reproducibility, if done	Applies to the reference standard as well as to the index test. Both should be reported/ adequately referenced. Report interrater and test-retest reliability of reference standard as established in the study being reported, rather than simply referring to other studies in which reproducibility has been established. The training that image readers receive should be carefully described. Studies in which the accuracy of "majority" judgments are reported should also report data for the minority judgments. Reports of the impact of training should clearly describe the characteristics of the sample used for training and whether it is representative of the group to which the test will be applied
Results		
Participants		
14	When study was performed, including beginning and end dates of recruitment	Pertinent particularly to longitudinal (delayed verification) studies, authors should report recruitment dates of the study (not to be confused with recruitment dates of the wider cohort study from which it might be drawn), and the beginning (first participant) and end (last participant) dates of the periods during which the index test(s) and reference standard were performed. Report the period for the index test and period for the reference standard separately if it is not clear
15	Clinical and demographic characteristics of the study population (at least information on age, sex, spectrum of presenting symptoms). See also item 18	Report key demographic variables: age, sex, and education. Report age distribution of sample in detail. Ethnicity and genetic factors (e.g., *APOE* genotype) may also be particularly important. The cognitive characteristics are covered in item 18
16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended). See also items 3–5	
Test results		
17	Time interval between the index tests and the reference standard, and any treatment administered in between	Specify the follow-up period for all subjects in relation to their outcomes. It should be specified whether participants had received any treatments that might affect disease progression
18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition	Include a description of the severity of the target condition at the time the index test is performed. Usually captured by a cognitive score and/or duration of symptoms. For delayed verification studies, report distribution of severity of disease and the degree of certainty (such as probable/possible) about the diagnosis at the time of case ascertainment. Report other diagnoses (not target condition). Report relationship of test to other diagnoses

Continued

Table 3 Continued

Section, topic, and item no.	STARD checklist item	Points of particular relevance to dementia
19	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	
20	Any adverse events from performing the index tests or the reference standard	Report all adverse events, even if unlikely to be related to the diagnostic test performed
Estimates		
21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals). See also item 12	
22	How indeterminate results, missing data, and outliers of the index tests were handled	
23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers, or centers, if done	
24	Estimates of test reproducibility, if done. See also item 13	
Discussion		
25	Discuss the clinical applicability of the study findings	Discuss differences in age and comorbidity between the study population and the patients typically seen in clinical practice. Discuss whether the reported data demonstrate "added" or "incremental" value of the index test over and above other routine diagnostic tests. Identify stage of development of the test (e.g., proof of concept; defining accuracy in a typical spectrum of patients). Discuss the further research needed to be done to make test applicable to population in whom likely to be applied in practice

Abbreviations: AD = Alzheimer disease; DLB = dementia with Lewy bodies; MeSH = Medical Subject Headings; STARD = Standards for Reporting of Diagnostic Accuracy.

the form of brief additions to the concise tabular format of STARD (see table 3; print version accessible from table e-1). For 18 of the 25 items, dementia-specific additions were deemed necessary. Also, dementia-specific examples of adequate reporting were derived from the literature and are available in table e-2.

There are 4 key areas central to effective evaluation of studies of diagnostic tests in dementia and cognitive impairment to which special attention should be given when reporting results of DTA studies in dementia and cognitive impairment:

1. Study population: Many DTA studies report on a highly selected population, e.g., from a tertiary memory clinic or a sample of convenience. Generalizability to the population with cognitive impairment at large, or even to other speciality clinics, may be questionable. The report should address whether that sample was representative in terms of spectrum of disorders, proportion of cases with the disease for which the index test is intended, and severity of cognitive impairment of the population in whom the test would be applied in practice. If not, then test accuracy may be over- or underestimated.
2. Reference standard: The current limitations in our reference standards in dementia are responsible for some of the inaccuracy and bias that bedevil studies of test accuracy. The 2 major classes of reference standard are (1) postmortem verification, and (2) progression from MCI to dementia due to AD or other conditions according to clinical criteria. Both fall short of the ideal and carry risks of disease misclassifications (bias by the reference standard). An inconsistently applied reference standard creates obvious difficulties in effectively evaluating the performance of a test across studies. Careful specification of the reference standard(s), its operationalization, and application are essential. If more than one reference standard is applied, the index test results should also be displayed by the different reference standards.
3. Circularity: "Incorporation bias," whereby the index test forms a part of the reference standard, is common in dementia diagnostic studies. This is inevitable given the composite nature of the reference standards and is a particular problem in the evaluation of neurocognitive tests. Incorporation bias is associated with a tendency to overestimate the value of a specific diagnostic test. The risk of such bias should be acknowledged and reported.

4. Reliability: Reporting on the test-retest reliability is important. Intra- and interobserver variability may have important effects on neurocognitive scales. For many of the CSF biomarkers, significant intraindividual variation can be found on repeated testing. In addition, there may be substantial interlaboratory variability.[32,33] Initiatives are underway to pinpoint and minimize causes of test variation, particularly for protein biomarkers.[34,35] Clear reporting of test-retest reliability should help to complement these efforts.

DISCUSSION It is striking that many of the causes of bias in studies of DTA are similar to those in clinical trials: population selection, blinding, missing data and dropouts, and unreliable outcomes. Other common causes of bias more specific to DTA studies are use of healthy controls and mixed reference standards.[15,36] Without full and transparent reporting, readers are unable to assess the validity of individual studies, and thus the overall body of evidence available for a particular test or biomarker. This increases the risk of misinterpretation and misuse of the test data.[37] High-quality reporting is of particular importance as patients increasingly present early with equivocal symptoms.[12] A diagnosis (or indeed, misdiagnosis) of such a disease has profound implications for patients and their families. Although no disease-modifying treatments or treatments for early-stage illness have reached clinical practice, it is imperative that, when they do, confidence in the diagnostic process is high. At present, the patchy quality of reporting damages the confidence with which findings from studies of DTA can be translated and applied to clinical practice.

The STARD guidelines are important in raising awareness of the key reporting issues in DTA studies.[25] However, despite widespread adoption by journals, our earlier work shows that standards of reporting in the dementia field are not uniformly high.[26] The STARDdem Consensus Initiative serves to raise awareness of the issues in test accuracy, and emphasizes those that are particularly important to dementia and cognitive impairment.

AUTHOR AFFILIATIONS

From the Cochrane Dementia and Cognitive Improvement Group (A.H.N.-S., J.M.M., N.S., S.M.), Department of Geratology, Nuffield Department of Clinical Neurosciences (G.W.), and Department of Psychiatry (R.M.), University of Oxford, UK; Department of Neurology, Academic Medical Centre (E.R.), and Department of Clinical Epidemiology and Biostatistics (P.M.M.B.), University of Amsterdam, the Netherlands; Centre for Mental Health (C.W.R.), Imperial College London, UK; Western Australian Centre for Health & Ageing–WACHA (L.F.), Western Australian Institute for Medical Research, University of Western Australia, Perth; Centre for Academic Mental Health (S.J.C.), School of Social and Community Medicine, University of Bristol; Institute of Public Health (D.D., C.B.), University of Cambridge; Institute of Cardiovascular and Medical Sciences (T.J.Q.), University of Glasgow, The New Lister Building, Glasgow; Peninsula Technology Assessment Group (PenTAG) (C.H.), University of Exeter, UK; Institute of Social and Preventive Medicine (A.W.S.R.), University of Bern, Switzerland; Sunnybrook Research Institute (S.B.), Sunnybrook Health Sciences Centre, Toronto, Canada; Institute of Neuroscience and Physiology (K.B., N.M., H.Z.), Department of Psychiatry and Neurochemistry, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden; Department of Neurology and Psychiatry (M.F.), University of Rome–Sapienza, Rome, Italy; Institute for Health & Aging, Department of Social and Behavioral Sciences (J.K.J.), and Departments of Neurology (J.K.J.) and Radiology and Biomedical Imaging (N.M.), University of California, San Francisco; Nursing Research Group (S.K.), Institute for Social Medicine and Epidemiology, University of Lübeck, Germany; Department of Psychiatry, Neurology, and Gerontology (L.S.S.), University of Southern California Keck School of Medicine; and King's College London (A.S.), Institute of Psychiatry, Department of Neuroimaging, London, UK.

AUTHOR CONTRIBUTIONS

Anna H. Noel-Storr: substantial contribution to conception and design of initiative and its methods; primarily responsible for formation and coordination of STARDdem Working Group; coordinated Web site design and content management during consensus period; collated consensus round comments; codrafted initial draft guidance; drafted first draft of manuscript; corevised subsequent drafts. Jenny M. McCleery: member of the STARDdem Working Group, contributed to conception and design of initiative and its methods; substantial input into codrafting of initial draft of guidance; substantial input into subsequent drafts based on consensus round feedback; substantial input into manuscript development. Edo Richard: member of the STARDdem Working Group; made substantial contributions during the redrafting of the guidance based on consensus round feedback; substantial input into the manuscript development. Craig W. Ritchie and Leon Flicker: member of the STARDdem Working Group; substantial role in conception of initiative; input into manuscript development. Sarah J. Cullum, Daniel Davis, Terence J. Quinn, Chris Hyde, Anne W.S. Rutjes, Nadja Smailagic, and Sue Marcus: member of the STARDdem Working Group; input into manuscript development. Sandra Black, Kaj Blennow, Carol Brayne, Mario Fiorivanti, Julene K. Johnson, Sascha Köpke, and Lon S. Schneider: contributed to the consensus round; contributed to the drafting of the manuscript. Andrew Simmons, Niklas Mattsson, Henrik Zetterberg, and Patrick M.M. Bossuyt: contributed during the consensus round; contributed to the drafting of the manuscript. Gordon Wilcock: Member of the STARDdem Working Group; chaired the STARDdem consensus meeting; contributed to the consensus round; contributed to the drafting of the manuscript. Rupert McShane: Coconvenor of the STARDdem Working Group; substantial contribution to conception and design of initiative and its methods; substantial input into codrafting of initial draft of guidance; substantial input into subsequent drafts based on consensus round feedback; substantial input into manuscript development.

ACKNOWLEDGMENT

STARDdem Initiative contributors (not including authors or members of the STARDdem Working Group) during the consensus process are listed on the *Neurology*® Web site at Neurology.org.

STUDY FUNDING

This project was funded by the National Institute for Health Research Cochrane Collaboration programme grant scheme: Programme Grant of Diagnostic Test Accuracy Reviews (project 10/4001/05). The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR, NHS, or the Department of Health.

DISCLOSURE

The authors report no disclosures relevant to the manuscript. Go to Neurology.org for full disclosures.

Received November 30, 2013. Accepted in final form April 7, 2014.

REFERENCES

1. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:263–269.
2. Sperling RA, Aisen PS, Beckett LA, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280–292.
3. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging and Alzheimer's Association workgroup. Alzheimers Dement 2011;7:270–279.
4. Dubois B, Feldman HH, Jacova C, et al. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol 2010;9:1118–1127.
5. FDA approves imaging drug Amyvid [online]. Available at: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm299678.htm. Accessed February 10, 2014.
6. European Medicines Agency. Summary of opinion: Amyvid, October 2012 [online]. Available at: http://www.ema.europa.eu/ema. Accessed February 10, 2014.
7. Johnson KA, Minoshima S, Bohnen NI, et al. Appropriate use criteria for amyloid PET: a report of the Amyloid Imaging Task Force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association. Alzheimers Dement 2013;9:e-1–e-16.
8. Isaac M, Vamvakas S, Abadie E, Jonsson B, Gispen C, Pani L. Qualification opinion of novel methodologies in the predementia stage of Alzheimer's disease: cerebrospinal-fluid related biomarkers for drugs affecting amyloid burden—regulatory considerations by European Medicines Agency focusing in improving benefit/risk in regulatory trials. Eur Neuropsychopharmacol 2011;21:781–788.
9. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984;141:1356–1364.
10. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53:695–699.
11. Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR. The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int J Geriatr Psychiatry 2006;21:1078–1085.
12. Goodyear-Smith F. Government's plans for universal health checks for people aged 40–75. BMJ 2013;347:f4788.
13. Cordell CB, Borson S, Boustani M, et al. Alzheimer's Association recommendations for operationalizing the detection of cognitive impairment during the Medicare Annual Wellness Visit in a primary care setting. Alzheimers Dement 2013;9:141–150.
14. Le Couteur DG, Doust J, Creasey H, Brayne C. Political drive to screen for pre-dementia: not evidence based and ignores the harms of diagnosis. BMJ 2013;347:f5125.
15. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469–476.
16. Smidt N, Rutjes AW, van der Windt DA, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235:347–353.
17. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006;67:792–797.
18. Young NS, Ioannidis JP, Al-Ubaydli O. Why current publication practices may distort science. PLoS Med 2008;5:e201.
19. Simera I, Moher D, Hoey J, Schulz KF, Altman DG. A catalogue of reporting guidelines for health research. Eur J Clin Invest 2010;40:35–53.
20. Simera I, Altman DG, Moher D, Schulz KF, Hoey J. Guidelines for reporting health research: the EQUATOR Network's survey of guideline authors. PLoS Med 2008;5:e139.
21. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–694.
22. Moher D, Hopewell S, Schulz F, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.
23. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Clin Chem 2003;49:1–6.
24. STARD. STARD statement [online]. Available at: www.stard-statement.org. Accessed February 10, 2014.
25. Ochodo EA, Bossuyt PM. Reporting the accuracy of diagnostic tests: the STARD Initiative 10 years on. Clin Chem 2013;59:917–919.
26. Noel-Storr AH, Flicker L, Ritchie CW, et al. Systematic review of the body of evidence for use of biomarkers in the diagnosis of dementia. Alzheimers Dement 2013;9:e96–e105.
27. EQUATOR Network. Enhancing the QUAlity and Transparency of Health Research [online]. Available at: www.equator-network.org. Accessed February 10, 2014.
28. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000;32:1008–1015.
29. National Institute for Health Research. Cochrane Dementia and Cognitive Improvement Group (CDCIG) [online]. Available at: http://dementia.cochrane.org. Accessed February 10, 2014.
30. Davis DHJ, Creavin ST, Noel-Storr AH, et al. Neuropsychological tests for the diagnosis of Alzheimer's disease dementia and other dementias: a generic protocol for cross-sectional and delayed-verification studies. Cochrane Database Syst Rev 2013;(3):CD010460.
31. STARDdem. Reporting standards for studies of diagnostic test accuracy in Alzheimer's disease and dementia: the STARDdem (STAndards for the Reporting of Diagnostic accuracy studies–dementia) Initiative [online]. Available at: http://www.starddem.org. Accessed February 10, 2014.
32. Verwey NA, van der Flier WM, Blennow K, et al. A worldwide multicentre comparison of assays for cerebrospinal fluid biomarkers in Alzheimer's disease. Ann Clin Biochem 2009;46:235–240.
33. Mattsson N, Andreasson U, Persson S, et al. CSF biomarker variability in the Alzheimer's Association Quality Control Program. Alzheimers Dement 2013;9:251–261.

34. Vanderstichele H, Bibl M, Engelborghs S, et al. Standardization of preanalytical aspects of cerebrospinal fluid biomarker testing for Alzheimer's disease diagnosis: a consensus paper from the Alzheimer's Biomarkers Standardization Initiative. Alzheimers Dement 2012;8:65–73.
35. Carrillo MC, Blennow K, Soares H, et al. Global standardization measurement of cerebral spinal fluid for Alzheimer's disease: an update from the Alzheimer's Association Global Biomarkers Consortium. Alzheimers Dement 2013;9:137–140.
36. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–1066.
37. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009; 374:86–89.