Takwoingi, Y. and Quinn, T. (2018) Review of Diagnostic Test Accuracy (DTA) studies in older people. Age and Ageing, 47(3), pp. 349-355.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/155648/

Deposited on: 18 January 2018

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Review of Research Methods

Diagnostic Test Accuracy Research in Older Adults

Abstract: Diagnostic test accuracy (DTA) describes a field of research that aims to assess how well a test is able to detect or exclude a condition of interest. Although geriatric medicine is not as reliant on investigations as other medical disciplines, almost all patient encounters with older adults will involve some form of diagnostic assessment. Thus, understanding the terminology and methods of DTA is essential for any clinician. In this review we use examples based around the diagnosis of dementia to highlight issues in DTA research. Some of these are generic to any DTA research and some are particularly pertinent to older adults. One can apply a test accuracy framework to a clinical question by defining four key components: the condition of interest; the index test(s) (i.e. the assessment(s) of interest); the reference standard (the best available method for assessing the condition of interest) and the population or healthcare setting in which testing takes place. Test accuracy is often described using complementary measures of sensitivity and specificity. However, many other metrics to describe test accuracy are available; in clinical practice predictive values may have greater utility. These and other descriptive statistics can be derived from a two by two table that cross-classifies the index test results with the reference standard results. Test performance and utility is not only determined by accuracy, other measures such as feasibility and acceptability should be considered and may be of particular importance when describing test performance in older adults with physical and cognitive impairments.

Keywords: accuracy; diagnosis; dementia, sensitivity; specificity; QUADAS, STARD

Is test accuracy relevant to older adult practice?

With so many research methods available to the older adult researcher, many of which have been reviewed in this Age and Ageing series [1,2], test accuracy research may not seem the most obviously relevant to our patient group. As a speciality, Geriatric Medicine prides itself on not being overly reliant on sophisticated imaging or laboratory tests. However, if we consider our clinical encounters we can see that concepts such as screening and diagnosis are central to geriatric practice. The resulting questions can be understood using a test accuracy research framework and so, by implication, an understanding of test accuracy research should be considered 'core business.(Figure 1)

The 'testing of tests' should not be the exclusive reserve of interested academics. Recent developments in imaging and 'omics-based technologies have resulted in an increasing clinical diagnostic toolkit. These new technologies offer exciting opportunities, but we should not assume that a new test is necessarily a better test. We must exercise the same rigour with test accuracy research as we would with a trial of a new treatment or device. The same argument holds for existing and established tests; it is sobering to see how many tests have become embedded within practice with little or no supporting test accuracy evidence.

The methodology that underpins test accuracy research is constantly evolving in-line with the developments in diagnostics and technology. In this review we will provide an introduction to the field with particular reference to the application of classical test accuracy in studies of older adults. For the reader wishing a more detailed discussion of the science and methodology of test accuracy, we have included key papers and a textbook as references.[3-6] The textbook addresses specific issues in dementia diagnostic test accuracy and performance of cognitive screening tests.[3]

Language of test accuracy research

Studies of diagnostic test accuracy (DTA) describe how well a test(s) correctly identifies or excludes a condition of interest (i.e. target condition). Test accuracy is only one phase in the multifaceted evaluation of a clinical test. The early phase of the process focusses on properties such as test reliability, precision, responsiveness and, following assessment of DTA, a later phase addresses the utility of the test in informing clinical decisions and ultimately improving patient outcomes.(Figure 2) Several authors have developed frameworks for the process of test evaluation, sometimes designed for specific test types such as laboratory tests [7] genetic tests [8], biomarkers [9] and imaging tests.[10]

Test performance depends on various factors such as characteristics of the test and its conduct (including expertise of assessors), purpose (e.g. diagnosis, screening, staging, disease surveillance, etc.), population and definition of the target condition.[11] Therefore, a clear definition of the intended use and role of a test for a specific population within the context of a clinical pathway is essential. Test accuracy is best understood by applying a framework that defines four key components: the target condition, population, index test(s), and reference standard. We will address each in turn and use the example of diagnosis of dementia to illustrate some of the challenges of DTA research in older adults.

Target condition: The target condition is the condition or clinical state that you wish to diagnose, exclude or differentiate. In some situations the target condition is obvious, for example using electrocardiograph or cardiac enzymes to diagnose cardiac ischaemia. However, the target condition should not be assumed and a clear, operationalised definition is mandatory. For example, in a chronic progressive syndrome such as dementia, we must be clear as to which form of dementia and which stage of the condition we are looking to diagnose. A test designed to assess for mild cognitive impairment (MCI) may have very different test properties if used to assess for frank dementia. Thus, we should be cautious of extrapolating DTA results from a narrow to a broader target condition. The popular Montreal Cognitive Assessment (MoCA) was designed and validated for MCI[12] but has now entered practice as a screen for dementia. Consequently the accuracy of MoCA as a dementia assessment differs from that reported in the original papers describing the test.[13]

Population: Measures of test accuracy are not usually transferable across different populations and settings due to changes in disease spectrum.[14] The spectrum of disease in a population will depend on prevalence, disease severity, clinical setting and prior testing. For example, a brief cognitive screening test will perform differently (and will have different preferred properties) when used in a specialist memory clinic compared to when used in a general practice consultation. In dementia DTA research, a particular issue is around the case-mix of the population and the potential severity of any cognitive syndrome. In a secondary care service, disease is likely to be advanced and differentiating those with and without disease may be relatively straightforward. If a test validated in secondary care is then used in primary care, where the proportion with disease is lower and disease is less advanced, test accuracy may differ considerably.

Index test: The index test describes the assessment(s) or tool(s) of interest. Index tests range from history taking and clinical symptoms to state of the art imaging and genomics. The index test may be more accurate, quicker, cheaper or less burdensome than the usual optimal test. If this is not the case one has to ask why the new test is required.

To enable analysis of test accuracy, the index test result is interpreted as positive or negative. As many biological states exist on a continuum, tests often give a range of values or scores. In this situation we have to define a threshold (cut-point) that categorises the index test result into positive and negative. Again, if we consider the MoCA, a range of scores are possible and a cut-point is used to determine 'test positive' and 'test negative' individuals.[13] While this dichotomisation is necessary for DTA analysis and allows ease of interpretation, important granularity can be missed. With a cognitive test it may be more useful to consider performance in each of the differing domains of memory, visuospatial, executive function tested rather than an aggregate score. Newer cognitive tests such as the Oxford Cognitive Screen actively discourage a reductionist pass/fail approach.[15]

Reference standard: The reference standard is the best available test for verifying the presence or absence of the target condition. This may be a single test or a combination of several tests and clinical information. Therefore, the description 'gold standard' is not always appropriate, as the reference standard may not be the optimal diagnostic approach. In dementia research many would consider the gold standard to be detailed neuropathology but clearly in large scale assessment of a cognitive screening test one could not expect every patient to offer neuropathological materials.

In practice, the reference standard is often expert clinical diagnosis but the synthesis of information and gestalt that informs clinical diagnosis has inherent limitations that need to be borne in mind when interpreting test accuracy. Even in expert hands, the diagnosis of dementia has a degree of inter-observer variability. If one then considers the differing classification systems available to diagnose dementia and dementia subtypes, the potential variation becomes even more pronounced.

Design of test accuracy studies

The usual design of a DTA study is cross-sectional. Participants representative of those in whom the test will be applied in practice are recruited sequentially or randomly, and all participants get the same index test and reference standard. For a new test, a case-control design is often used, i.e., the test is assessed using one group of participants known to have the target condition and another group without the target condition (two-gate design).[16] This design is a useful first step but is rarely the 'final word' on accuracy because such designs tend to inflate test accuracy by including phenotypic extremes.[16] Ideally, promising results from a case-control DTA study should be followed by a prospective study of participants suspected of having the condition. If the target condition is rare, researchers sometimes 'enrich' a population with additional cases. For example, a test designed to specifically look for the frontotemporal form of dementia may include all referrals to a memory clinic and additional cases already diagnosed.

Most published DTA studies assess the accuracy of a single index test. This is often not the clinical question of greatest relevance. For many conditions, particularly dementia assessment, there is more than one potential index test and the clinician will want to know which test is more accurate for a certain population. For example, how does MoCA compare with Folstein's Mini-Mental State Examination (MMSE)[17] for detecting MCI in community-dwelling older adults? Ideally, the relevant index tests should be compared 'head-to-head' in the same study population. This can be achieved by performing all the index tests on each participant (paired or within-subject design), or by randomising participants to a particular index test.[18,19] In both designs, all participants get the reference standard. Depending on the nature of the tests, the paired design may increase the burden of testing and so may not be ethically feasible, while the randomised design potentially requires a larger sample size.

Quantifying test accuracy

Ideally, an index test should discriminate perfectly between those with and those without the target condition, i.e. no false negative or false positive test errors. In clinical practice, this is rare. By cross-classifying index test and reference standard results, we create a two by two table of the number of

true positives, false positives, true negatives and false negatives. Various metrics for quantifying test accuracy can be computed using the data in this table.(Figure 3) When reporting test accuracy it is good practice to present the table or give sufficient information to allow the reader to reconstruct the table.

Sensitivity, specificity, and threshold effect: Sensitivity and specificity are the most commonly used test accuracy measures.[20] Sensitivity, also known as the true positive rate, is the proportion of those with the target condition that are correctly identified by the index test as 'cases'. Specificity, also known as the true negative rate, is defined as the proportion of those without the target condition that are correctly identified by the test as 'non-cases'.

There is a negative correlation between sensitivity and specificity induced by varying test threshold. Consider MMSE for assessment of dementia where the traditional cut-off is 24. If we use a higher cut-point, we will detect more people with dementia (higher sensitivity) but at the expense of labelling more people without dementia as having the condition (lower specificity). The receiver operating characteristic (ROC) plot is a graphical illustration of this trade-off between sensitivity and specificity.(Figure 4) Traditionally, sensitivity is plotted against 1-specificity for a range of thresholds. The position of the resulting ROC curve indicates the accuracy of the test; the higher the accuracy of the test, the closer the curve is to the upper left hand corner of the plot where sensitivity and specificity are both 100%. The area under the ROC curve (AUC) is often used as a global measure of test accuracy, with higher values signifying greater test accuracy. As the AUC is a single measure, when used in isolation its clinical utility is limited because it does not provide any information on how patients are misclassified (i.e. numbers of false positive and negative). The ROC plot can be used to compare the accuracy of different tests within a study.

Positive and negative predictive values: In clinical practice our interest is usually whether the test result helps classify the patient. The positive predictive value (PPV) is the proportion of those with positive test results who truly have the disease while the negative predictive value (NPV) is the proportion of those with negative test results who truly do not have the disease. As stated earlier, test performance is susceptible to disease spectrum, but predictive values in particular are directly affected by prevalence. Therefore, assessment of predictive values is inappropriate for case-control studies and enriched samples because prevalence in the sample is artificial.

Applying test accuracy in practice

For the clinician faced with a variety of differing test strategies[21], a common question is 'what values of sensitivity and specificity would suggest a good test'. There is no correct answer as the preferred trade-off between sensitivity and specificity is dependent on context. One must consider the purpose of testing and the implications of a false positive and false negative result. We can illustrate this point with a topical example. There has been considerable recent interest in CSF based dementia biomarkers. Abnormal levels of amyloid or tau proteins are associated with future risk of Alzheimer's disease dementia. However, when CSF biomarker properties are described using two by

two table based metrics, sensitivity and specificity of the biomarkers for diagnosis of undifferentiated dementia are not perfect. The implications of this are worth considering. If the test gives a false negative result in a middle aged person with early stage Alzheimer's disease, then the person will be misdiagnosed as normal. With no curative treatment for preclinical (or clinical) dementia this mislabelling has limited clinical implications and the disease will eventually be diagnosed when symptoms become apparent. However, consider another person without dementia who receives a false positive result. They will be misdiagnosed as having a progressive neurodegenerative condition with likely substantial negative effects on psychological health and implications for insurance, employment etc. Thus, one could argue that for the dementia screening scenario we would want the test to be highly specific and would accept a poorer sensitivity.

For a new test, the question may be 'what is the added value of the new test beyond what is already known from previous tests?' For example, in a study looking at CSF dementia biomarkers, the authors found reasonable test accuracy of these biomarkers, but when considered in the context of standard memory testing, there was little additional value of the novel, expensive invasive tests.[23] Quantifying the added value of a new test over an existing test paradigm is complex and, as alluded to in the preceding section, needs to consider the relative consequences of changes in true positive and true negative rates.[22]

Moving beyond sensitivity and specificity

Test accuracy is only part of the clinical evaluation process. The clinical value of a test depends on whether the information provided leads to improved patient outcomes.[24] Accuracy is not synonymous with clinical effectiveness and there are many plausible reasons why a new test with greater accuracy may not result in improved patient outcomes. The framework by Ferrante di Ruffano et al describes the mechanisms that commonly affect health outcomes including feasibility, acceptability, interpretability of test results and clinician confidence in the test.[25] Thus, once potentially favourable test accuracy is demonstrated this should be followed by 'real world' assessments that describe the test-treatment-outcomes pathway. Ultimately, the most important measure is whether the use of the test improves clinical outcomes (clinical utility studies). The test-treatment randomised controlled trial (RCT) is regarded as the ideal study design for evaluating clinical utility. In such RCTs, patients are randomised between new and existing tests, followed by appropriate management or intervention based on test results, and finally patient outcomes are measured and assessed. However, the cost and duration of these RCTs often make them unrealistic.[24]

In the context of testing in the older adult, test accuracy needs to be considered alongside other properties including feasibility and acceptability. Greater test accuracy may come at the cost of increased test administration time, more invasive testing and greater test burden for both the patient and the assessor. An accurate test that can only be completed by a small proportion of the intended population has limited value. This is an issue for all healthcare settings but is likely to be particularly pertinent when assessing frail older adults. To continue our theme of dementia assessment, we can consider multi-domain neuropsychological testing (neuropsychological battery [NPB]). In many dementia texts, NPB is considered the ideal assessment strategy, in fact some studies use NPB as the 'gold standard' for diagnosing dementia. However, the lengthy, detailed

testing required may not be tolerated and this non-completion is likely to be over-represented in those with underlying cognitive problems. The potential bias caused by this partial verification should not be under estimated and the way missing values of index test and reference standard are handled in DTA analysis can lead to very different estimates of test accuracy.[26] Some authors have argued that the classical two by two table should be supplemented by additional cells describing the numbers who are 'untestable'.

Assessing reporting and methodological quality

As this brief review has highlighted, test accuracy research is challenging. Deficiencies in reporting have been recognised as a particular problem for DTA research. In response, the Standards for Reporting Diagnostic Accuracy (STARD) working group created best practice guidance.[4] There is evidence that reporting standards for DTA research are now improving [27] and this is at least in part driven by journals (including Age and Ageing) mandating that authors of DTA research follow STARD guidance. Test accuracy is a fast moving field and an update of STARD guidance was published in 2015, along with guidance on DTA reporting in abstracts and specific DTA guidance for dementia.[28] Comprehensive, transparent reporting will not save a methodologically flawed study and as well as reporting, guidance around the design and conduct of DTA research is also useful. The Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS) and more recently the updated QUADAS-2 tool offers a framework for assessing DTA studies in terms of patient selection, index test, reference standard, and participant flow and timing of assessments.[6] A more detailed discussion of QUADAS is available in an earlier Age and Ageing review.[2] STARD and QUADAS-2 can guide critical appraisal of DTA work and so have particular value in systematic review and meta-analysis of DTA.[29,30] However, even if not considering formal systematic synthesis of DTA research, we would encourage anyone considering running, peer-reviewing or simply reading the report of a test accuracy study to make use of QUADAS-2 and STARD resources.

References

1. Harrison JK, Reid J, Quinn TJ, Shenkin SD. Using quality assessment tools to critically appraise ageing research:A guide for clinicians. Age Ageing.2017;46:359-365

2. Witham MD, Stott DJ. Conducting and reporting trials for older people. Age Ageing.2017; 46:889-894.

3. Quinn T, Takwoingi Y. Assessment of the utility of cognitive screening instruments. In: Larner AJ (editor), Cognitive Screening Instruments: A Practical Approach (2nd Edition). Springer International Publishing; 2017.

4. 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. BMJ 2003;326:41-4.

5. Ritchie CW, Terrera GM, Quinn TJ. Dementia trials and dementia tribulations. Methodological and analytical challenges in dementia research. Alzheimers Res Ther. 2015;7:31

6. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.

7. Silverstein MD, Boland BJ. Conceptual framework for evaluating laboratory tests: case-finding in ambulatory patients. Clin Chem 1994;40:1621-7.

8. Haddow JE, Palomaki GE. ACCE: A model process for evaluating data on emerging genetic tests. Oxford: Oxford University Press; 2003.

9. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta 2014;427:49-57.

10. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987;60:1071-81.

11. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.

12. 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;53:695-9.

13. Davis DH, Creavin ST, Yip JL, Noel-Storr AH, Brayne C, Cullum S. Montreal Cognitive Assessment for the diagnosis of Alzheimer's disease and other dementias. Cochrane Database Syst Rev. 2015;10:CD010775.

14. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669-71.

15. Demeyere N, Riddoch MJ, Slavkova ED, Bickerton WL, Humphreys GW. The Oxford Cognitive Screen:validation of a stroke-specific short cognitive screening tool.Psychol Assess.2015;27:883-94.

16. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335-41.

17. O'Bryant SE, Humphreys JD, Smith GE, Ivnik RJ, Graff-Radford NR, Petersen RC, et al. Detecting dementia with the mini-mental state examination in highly educated individuals. Arch Neurol 2008;65:963-7

18. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-92.

19. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013;158:544-54.

20. Moons KG, Harrell FE. Sensitivity and specificity should be de-emphasised in diagnostic accuracy studies. Acad Radiol. 2003;10:670-2.

21. Harrison JK, Noel-Storr AH, Demeyere N, reynish EL, Quinn TJ. Outcome measures in a decade of dementia and mild cognitive imapirment trials.Alz Res & Therapy.2016;8:48

22. Pepe MS. The statistical evaluation of medical tests for classification and prediction. First edition. Oxford, UK: Oxford University Press; 2003.

23. Richard E, Schmand BA, Eikelenboom P, Van Gool WA, The Alzheimer's Disease Neuroimaging Initiative. MRI and cerebrospinal fluid biomarkers for predicting progression to Alzheimer's disease in patients with mild cognitive impairment: a diagnostic accuracy study. BMJ Open. 2013;3:e002541.

24. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850-5.

25. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012;344:e686.

26. Lees RA, Hendry K, Broomfield N, Stott DJ, Larner AJ, Quinn TJ.Cognitive assessment in stroke:feasibility and test accuracy using differing approaches to scoring of incomplete items.Int J Geriatr Psychiatry.2016;32:1072-78.

27. Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting Diagnostic Accuracy Studies: Some Improvements after 10 Years of STARD. Radiology 2015;274:781-9.

28. Noel-Storr AH, McCleery JM, Richard E et al. Reporting standards for studies of diagnostic test accuracy in dementia: the STARDdem Initiative. Neurology. 2014;83:364-73.

29. Harrison JK, Fearon P, Noel-Storr AH, McShane R, Stott DJ, Quinn TJ. IQCODE for diagnosis dementia in secondary care settings. Cochrane Database Syst Rev. 2015;3:CD010772.

30. Davis DH, Creavin ST, Noel-Storr A 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.

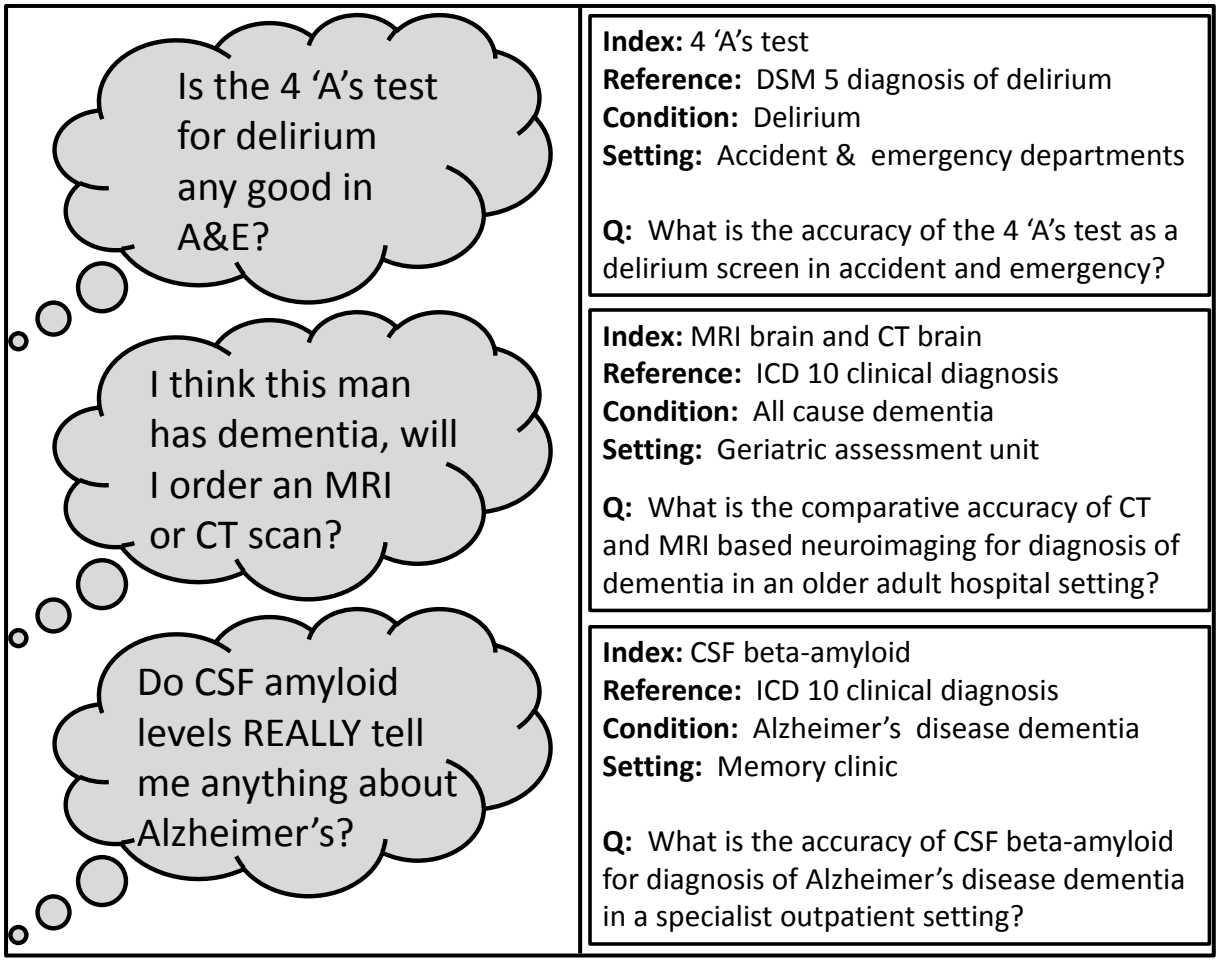

Figure 1. Diagnostic test accuracy in geriatric medicine.

A&E, accident and emergency; CSF, cerebrospinal fluid; CT, computed tomography; DSM 5, diagnostic and statistical manual of mental disorders fifth edition; ICD10, international classification of diseases tenth revision; MRI, magnetic resonance imaging.

The cartoon illustrates how clinical questions can be formatted as hypotheses suitable for diagnostic test accuracy research.

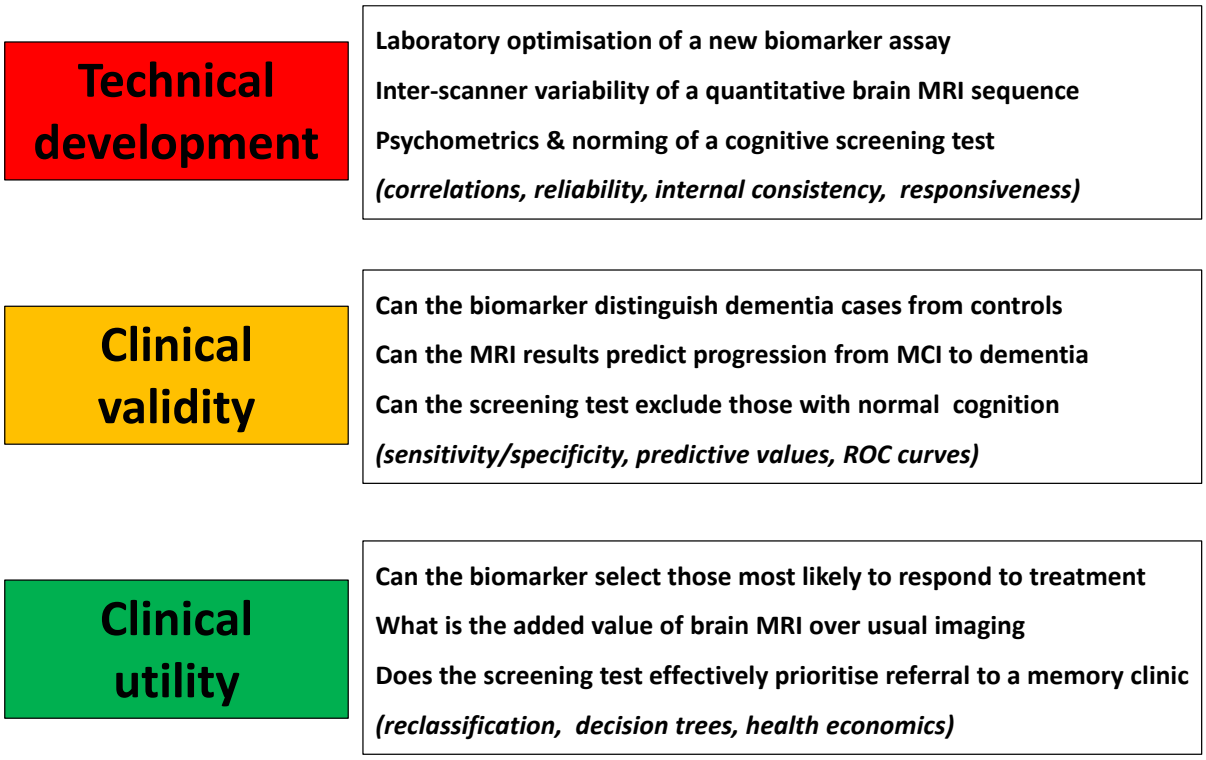

Figure 2. Potential phases of test development and assessment.

MRI, magnetic resonance imaging; ROC, receiver operating characteristic.

	Reference standard present (dementia, ICD-10 criteria)	**Reference standard absent** (dementia, ICD-10 criteria)	
Index test (eg MMSE) **positive**	True positives (A)	False positives (B)	PPV A ÷ (A + B)
Index test (eg MMSE) **negative**	False negatives (C)	True negatives (D)	NPV D ÷ (C + D)
	Sensitivity A ÷ (A + C)	Specificity D ÷ (B + D)	

Other paired measures of accuracy:

Positive likelihood ratio (LR+): sensitivity ÷ (1 – specificity)
Negative likelihood ratio (LR-): (1 – sensitivity) ÷ specificity

False alarm rate: 1 - PPV
False reassurance rate: 1 - NPV

Single measures of accuracy

Diagnostic odds ratio (DOR): (A x D) ÷ (B x C)
Youden index: sensitivity + specificity – 1
Overall accuracy: (A + D) ÷ (A + B + C + D)

Figure 3. The classical two by two test accuracy table and metrics that can be derived from this table.

MMSE, Mini-Mental State Examination.

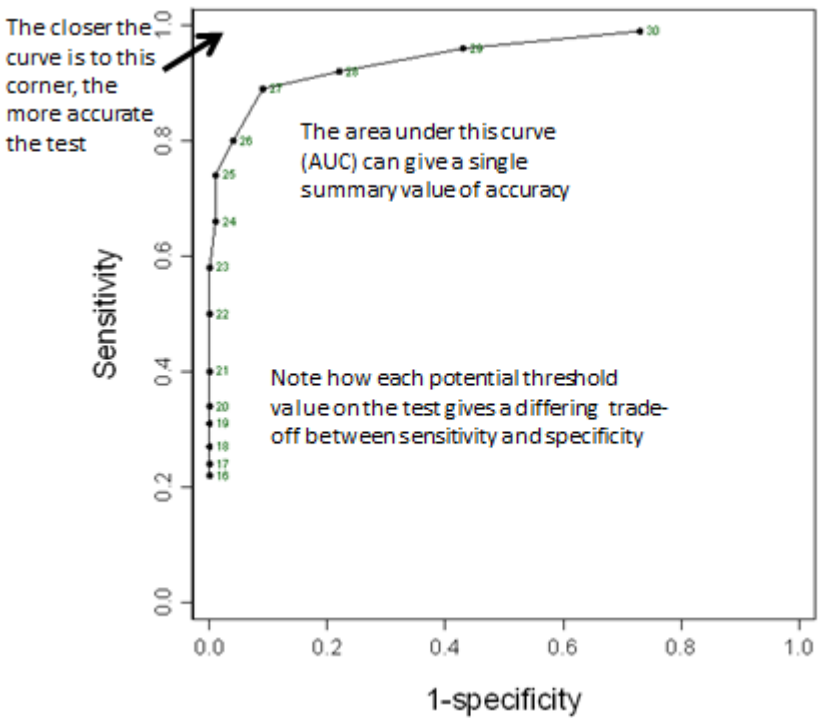

Figure 4. Receiver operating characteristic curve of Mini-Mental State Examination for detecting dementia

The numbers within the figure indicate the cut-off score for each pair of sensitivity and specificity. *Figure authors' own with data based on [17]*