Diagnostic test accuracy of remote, multidomain cognitive assessment (telephone and video call) for dementia

Quinn, T.J.; Elliot, E.; Hietamies, T.M.; Martinez, G.; Tieges, Z.; Mc Ardle, R.

Published in:
Cochrane Database of Systematic Reviews

DOI:
10.1002/14651858.CD013724

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Cochrane Library

Cochrane Database of Systematic Reviews

Diagnostic test accuracy of remote, multidomain cognitive assessment (telephone and video call) for dementia (Protocol)

Quinn TJ, Elliott E, Hietamies TM, Martínez G, Tieges Z, Mc Ardle R

Quinn TJ, Elliott E, Hietamies TM, Martínez G, Tieges Z, Mc Ardle R.
Diagnostic test accuracy of remote, multidomain cognitive assessment (telephone and video call) for dementia (Protocol).
Cochrane Database of Systematic Reviews 2020, Issue 9. Art. No.: CD013724.
DOI: 10.1002/14651858.CD013724.

www.cochranelibrary.com

TABLE OF CONTENTS

[Diagnostic Test Accuracy Protocol]

Diagnostic test accuracy of remote, multidomain cognitive assessment (telephone and video call) for dementia

Terry J Quinn[1], Emma Elliott[1], Tuuli M Hietamies[1], Gabriel Martínez[2], Zoë Tieges[3], Riona Mc Ardle[4]

[1]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. [2]Faculty of Medicine and Dentistry, Universidad de Antofagasta, Antofagasta, Chile. [3]Department of Geriatric Medicine, Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK. [4]Translational and Clinical Research Institute, Newcastle University, Newcastle, UK

Contact address: Terry J Quinn, Terry.Quinn@glasgow.ac.uk, tjq1t@clinmed.gla.ac.uk.

Editorial group: Cochrane Dementia and Cognitive Improvement Group.
Publication status and date: New, published in Issue 9, 2020.

Citation: Quinn TJ, Elliott E, Hietamies TM, Martínez G, Tieges Z, Mc Ardle R. Diagnostic test accuracy of remote, multidomain cognitive assessment (telephone and video call) for dementia (Protocol). *Cochrane Database of Systematic Reviews* 2020, Issue 9. Art. No.: CD013724. DOI: 10.1002/14651858.CD013724.

ABSTRACT

Objectives

This is a protocol for a Cochrane Review (diagnostic). The objectives are as follows:

Our main objective is to assess the test accuracy of any multidomain cognitive test delivered remotely for the diagnosis of any form of dementia.

Our review will not be limited to a particular healthcare setting or a particular threshold score of index test.

Secondary objectives

- To describe the degree of agreement between a remotely delivered cognitive test and the same or a closely related test delivered in-person when neither is assessed against a clinical dementia reference standard.
- To identify the quality and quantity of the research evidence describing test accuracy of remote testing.
- To identify sources of heterogeneity in the test accuracy described.
- To identify gaps in the evidence where further research is required.

Investigation of sources of heterogeneity

Heterogeneity is often seen in clinical test accuracy reviews (Deeks 2001). Important potential sources of heterogeneity will include the case-mix of the population being assessed; clinical setting; person performing the assessment; platform used to administer the test (e.g. standard telephone versus video call); threshold scores used to define test positivity; and the quality of the included papers. We will collect data on all of these factors, and if data allow will explore their effect using subgroup and sensitivity analyses as appropriate.

BACKGROUND

Assessment of cognition using a multidomain test can serve many important purposes (Lin 2013). In clinical practice, cognitive testing may form part of the assessment of the person with a suspected cognitive syndrome, or the testing may be used as an initial triage tool to identify those who need more specialist input. In research, cognitive testing may be used to identify potential participants for a study or as an assessment of treatment effect in a clinical trial.

There are many multidomain cognitive assessment tools available to the clinician (Harrison 2016). Indeed, in some areas there are almost as many assessment tools as there are research studies (Lees 2012). It is important to distinguish the short screening tests that will be the focus of this review from more detailed assessments that attempt a diagnostic formulation. Although there is no consensus on the optimal cognitive assessment, certain tests have greater visibility and traction in research and practice. An important factor to consider when choosing a cognitive test is the test's accuracy for the detection of the condition of interest, for example the accuracy of a screening test for detection of dementia. In the Cochrane Dementia and Cognitive Improvement Group (CDCIG), we have reviewed the literature and summarised the accuracy of many of the commonly used cognitive screening tests, including Folstein's Mini-Mental State Examination (Creavin 2016), Montreal Cognitive Assessment (Davis 2015), and Addenbrooke's Cognitive Examination (Beishon 2019).

To date, our suite of diagnostic test accuracy (DTA) reviews have been limited to in person, face-to-face assessment, as this is the favoured clinical reference standard and would be usual practice in most services (Davis 2013). The current coronavirus pandemic has caused a fundamental change in practice that no one had anticipated. The emergency restrictions on movement and social contact necessitated by the viral pandemic limit the opportunity for in-person assessment. Clinical services and research teams have responded, and increasingly consultations are being performed remotely. Various cognitive screening tests designed for administration over the telephone or via video-call are available and could be well suited to the current situation (Elliott 2020). However, the diagnostic accuracy of these tools should not be assumed, and we felt it was necessary to collate, appraise, and present estimates of accuracy for papers describing the use of remote cognitive testing. Remote testing can be performed using the telephone, but increasing availability of audio-visual technologies also allows for assessment using video-based calls or other telehealth approaches. We are interested in both telephone and video-based assessments, but will consider these technologies separately.

Beyond the pandemic situation, there are other circumstances in which remotely administered cognitive tests could be useful. Many practitioners in remote and rural areas are already familiar with using them for clinical purposes (Barth 2018). In the research context, such tests may be the only feasible way to include cognitive outcome measures in large pragmatic trials or observational studies (Ritchie 2015).

Target condition being diagnosed

The condition of interest for this review is clinical dementia. We recognise that cognitive testing may be used to inform the diagnosis of other cognitive syndromes such as mild cognitive impairment (MCI), but the condition of greatest relevance is invariably clinical dementia. The dementia diagnosis is operationalised in various classification systems, such as the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) and the World Health Organization's International Classification of Diseases (ICD) (APA 2013; WHO 2010). Although there are some differences between these classifications (e.g. the most recent DSM guidance suggests use of the terms 'Major' and 'Mild Neurocognitive Disorder' rather than dementia and MCI), they all describe dementia as a progressive, irreversible condition characterised by impairments in multiple cognitive domains sufficient to cause problems in activities of daily living. Additional classifications exist to describe pathological dementia subtypes, for example Alzheimer's disease or vascular dementia.

Dementia test accuracy studies have traditionally used the paradigm of assessing a test of interest against a clinical dementia reference standard (Takwoingi 2018). For this particular review, with its focus on remote testing, we anticipate an alternative but equally important study design – where the remote test is compared against the usual, in-person administration of the same test. Although ultimately the cognitive testing is being performed to assess for a cognitive syndrome such as dementia, these papers may not necessarily include any dementia diagnosis data. Even without the dementia diagnosis, such papers can offer useful insights into the properties of remote testing compared to usual face-to-face practice. So, whilst such analyses are not diagnostic test accuracy, we will also review those papers that use a remote versus face-to-face test comparison approach.

Index test(s)

Our index test of interest is any multidomain cognitive assessment tool that is administered remotely, for example over the telephone or via video call. We will consider these two technologies separately.

We suspect that in most instances the test will be a variation of standard face-to-face cognitive assessment, although content may need to be modified for remote delivery. We are interested in real-time assessment that involves someone administering and scoring the test. Our remit will thus not extend to computer-based cognitive gaming that purports to offer an assessment, or online cognitive questionnaires.

We anticipate that the review will include several different tests. If the data allow, there may be an opportunity for comparative analyses of the accuracy of the various tests.

Clinical pathway

For consistency with other reviews and in keeping with usual methods, we use the phrase 'diagnostic test accuracy' (Davis 2013). However, we recognise that our index tests of interest are screening in nature and not sufficient to make a diagnosis on their own. In clinical practice the multidomain test is often the first step in a detailed, multidisciplinary assessment that may also be informed by assessments of function, collateral history, and radiological and laboratory testing (Noel-Storr 2012).

The remote tests of interest are not exclusive to a particular healthcare setting. Brief cognitive screening may be performed in primary care to inform the need for onward referral to specialist

services. In secondary care clinic settings, cognitive screening may be performed as part of a diagnostic work-up, or to monitor disease progression. All of these test scenarios could plausibly be performed remotely, and indeed the current viral pandemic is mandating this approach to testing. Prevalence of dementia will vary by setting, and if data allow we will explore this as part of our investigation of heterogeneity.

For inpatient secondary care services, cognitive screening often forms part of the initial assessment of admissions to the general hospital. Many countries recommend early cognitive screening of certain groups such as unscheduled older adult admissions or stroke survivors (Robinson 2015). In this situation, the purpose of testing is to identify a cognitive baseline and assess for the syndrome of delirium. This pathway is likely to remain an in-person assessment by the admitting team, and so is unlikely to be included in this review, but there will be no exclusions based on setting or purpose.

Alternative test(s)

Another alternative to in-person cognitive assessment is a questionnaire-based test (Harrison 2015), delivered either by post on using online platforms. We will not consider questionnaire-based approaches in this review, but hope to produce a separate review on self-complete questionnaire-based assessments.

Rationale

The arguments for timely diagnosis of dementia have been made by various professional societies and will not be rehearsed again here (Robinson 2015). Suffice it to say, cognitive testing is fundamental to the assessment of the person with a suspect cognitive problem. Our motivation for this review is in response to the current global viral pandemic, where best practice of in-person cognitive testing is often not safe or possible. Clinical and research teams are having to rapidly adapt to new ways of working, and we hope to provide an evidence base to guide choice of testing.

OBJECTIVES

Our main objective is to assess the test accuracy of any multidomain cognitive test delivered remotely for the diagnosis of any form of dementia.

Our review will not be limited to a particular healthcare setting or a particular threshold score of index test.

Secondary objectives

- To describe the degree of agreement between a remotely delivered cognitive test and the same or a closely related test delivered in-person when neither is assessed against a clinical dementia reference standard.
- To identify the quality and quantity of the research evidence describing test accuracy of remote testing.
- To identify sources of heterogeneity in the test accuracy described.
- To identify gaps in the evidence where further research is required.

Investigation of sources of heterogeneity

Heterogeneity is often seen in clinical test accuracy reviews (Deeks 2001). Important potential sources of heterogeneity will include the case-mix of the population being assessed; clinical setting; person performing the assessment; platform used to administer the test (e.g. standard telephone versus video call); threshold scores used to define test positivity; and the quality of the included papers. We will collect data on all of these factors, and if data allow will explore their effect using subgroup and sensitivity analyses as appropriate.

METHODS

Criteria for considering studies for this review

Types of studies

Our primary interest will be cross-sectional studies, where the index test(s) are administered alongside reference standard clinical assessment.

As a secondary, exploratory analysis we also propose a review of those papers where a remote assessment is compared to the equivalent in-person test. These papers may or may not also have information on reference standard dementia assessment. Where data are available that are limited to comparison of remote and in-person testing, we will include these studies but the proposed analysis will differ, recognising that these studies are assessing correlation/agreement rather than diagnostics.

We will exclude case-control studies due to the inherent risk of bias and inability to use these data to make any inferences about population predictive value. For the same reasons, we will exclude studies that use an enriched sample, for example studies that only include participants who have a certain screening test score, or where the reference standard assessment is limited to participants with a particular cognitive profile.

Studies where the index test is compared against future development of a cognitive syndrome (delayed-verification studies) require a differing review approach compared to the traditional cross-sectional test accuracy study. We will not consider delayed-verification studies in this review. We will not include any study where the index and reference standard are administered with more than one month between them, as such studies should be considered prognostic. Studies where the delay between index and reference is shorter are potentially eligible, and we will consider the effect of the delay as part of the 'Risk of bias' assessment.

We will exclude studies with a small number of cases (fewer than 10), as these studies are unlikely to meaningfully add to our understanding of test accuracy and are prone to various selection biases.

Participants

Our population of interest is any adult (age over 18 years) requiring cognitive testing.

We will not include studies exclusively comprised of cognitively normal participants that are used to create normative values of tests. For studies that compare a remote index to the equivalent face-to-face version, formal cognitive status may be unknown. We

will include these studies but will not combine them with studies that assess accuracy against a clinical dementia reference standard.

We will not exclude papers on the basis of a selected population, but where the population are not predominantly an older adult group (e.g. studies in traumatic brain injury or in stroke), we will note this and where possible explore case-mix as a source of heterogeneity.

We will include studies conducted in any healthcare setting, and explore setting as a source of heterogeneity where possible.

Index tests

Studies must include, not necessarily exclusively, a remote cognitive assessment.

Remote testing involves real-time assessment by a tester in a different location to the person being tested. Any platform that allows remote testing will be included, and we anticipate studies using traditional telephone, smartphone, videoconferencing.

Assessments must be multidomain, as these are the tests used in clinical practice. Tests of single cognitive domains such as memory only or attention only will therefore not be included.

Included tests may be modifications of existing in-person tests or bespoke assessments designed for remote use.

The assessment should be performed remotely, so we will not include studies where, for example, the script of a telephone interview is used in a face-to-face assessment.

Included tests should directly assess the person of interest. We will thus not include informant-based tests such as AD-8 (Hendry 2019) or IQCODE (Harrison 2015). If a test includes both informant responses and direct testing, and these data are available separately, we will include the direct test data. Where tests have contingent scoring, for example if a person scores above a certain value then further testing is performed, we will assess suitability for inclusion case by case, but will not pool these data with other screening tests.

Where we are comparing a remotely delivered index test to an in-person equivalent, we will include those tests from which the remote assessment was derived. For example, the various iterations of the Telephone Interview for Cognitive Status were designed to emulate the Folstein Mini-Mental State Examination (MMSE) (Brandt 1988), and so we will consider MMSE as a suitable reference for comparison. For this second objective, we will compare each index test to the same test administered face-to-face. If an index test was modified from a face-to-face parent test specifically for remote administration, then we will compare it with that parent test.

We will not limit the review to a particular remote test strategy, and anticipate including multiple index tests. If data allow, there may be an opportunity to perform indirect comparisons of estimates of accuracy of various remote tests, but these analyses will be exploratory rather than definitive (Owen 2018).

Our focus for this review is accuracy of testing. We will formally assess whether remote testing is feasible, acceptable, or suitable for the populations being tested, although some of these factors may be relevant to our internal and external validity assessments.

Target conditions

We will consider papers reporting any clinical diagnosis of dementia. Dementia diagnosis may be undifferentiated, or a particular subtype may be specified. Classifying dementia by subtype is not required for inclusion, but where available these data will be recorded.

Reference standards

Our reference standard will be a clinical diagnosis of dementia. Within the clinical diagnosis rubric, we will include all-cause (unspecified) dementia, using any recognised diagnostic criteria, for example ICD-10 or DSM-IV. For the purposes of this review, and in keeping with other Cochrane Diagnostic Test Accuracy Reviews, we will consider structured interview assessments such as Clinical Dementia Rating (CDR) as diagnostic (Davis 2013). Dementia diagnoses may specify a pathological subtype, and we will include all dementia subtypes in this review. We will not include the cognitive syndrome of delirium in the review (Hendry 2016), and will assume that in the process of making the clinical dementia diagnosis, any reversible causes of cognitive impairment would have been excluded.

Studies that make a postmortem diagnosis or base diagnosis on imaging or other biomarkers without corresponding comprehensive clinical assessment will not be eligible.

We will not set any limits in relation to severity or stage of dementia. If data are available from sufficient studies, we will explore the severity of dementia as a potential source of heterogeneity.

As a secondary, exploratory analysis, we also propose a review of those papers where a remote assessment is compared to the equivalent in-person test. These papers may or may not also have information on reference standard dementia assessment. Where data limited to comparison of remote and in-person testing are available, we will not use the index versus reference standard paradigm, but will describe agreement or correlation between the tests.

Search methods for identification of studies

Electronic searches

We will search MEDLINE (Ovid SP), Embase (Ovid SP), Science Citation Index (ISI Web of Knowledge), PsycINFO (Ovid SP), LILACS (Latin American and Caribbean Health Science Information database) (BIREME), and US National Institutes of Health Ongoing Trials Register ClinicalTrials.gov (www.clinicaltrials.gov/) databases. Each source will be searched from inception to the present. See Appendix 1 for a proposed draft strategy to be run in MEDLINE (OvidSP). We will design similarly structured search strategies using search terms appropriate for each database. We will use controlled vocabulary such as MeSH terms and EMTREE where appropriate. In the searches developed, we will make no attempt to restrict studies on the basis of sampling frame or setting. This approach is intended to maximise sensitivity and allow for inclusion on the basis of population-based sampling to be assessed at screening (see Selection of studies). We will not use search filters (collections of terms aimed at reducing the number needed to screen) as an overall limiter because those that are published have not proved sensitive enough (Whiting 2008). We will not apply any language restriction to the electronic searches, using translation services as needed. We will search the ALOIS, the

CDCIG Specialized Register, which includes both intervention and diagnostic test accuracy studies in dementia. We will not search the grey literature. We will perform forward and backward searching of included citations.

Searching other resources

NA

Data collection and analysis

Selection of studies

Following searching, titles from various databases will be collated in Covidence software (Covidence 2020). A single review author will perform a 'first pass' review, removing clearly irrelevant titles, then a minimum of two review authors will independently assess studies for eligibility.

For consistency with our other DTA titles, we will adopt a hierarchical approach to exclusion, first excluding on the basis of index test and reference standard, and then on the basis of study methods (case-control, size), and then on the basis of any other reason.

Where a potentially relevant paper is missing data needed for analyses, we will contact the primary author by email twice. If the authors do not respond, or the relevant data are not available, we will not include data from this study and label as 'data not suitable for analysis'. If the same dataset is presented in more than one paper, we will include the primary paper but make reference to other papers if they contain relevant additional information.

Where studies are described in abstract form, we will contact the lead author(s) to ask if the full paper is published. We will limit the studies included to those published in peer-reviewed scientific journals.

We will detail the study selection process in a PRISMA flow diagram.

Data extraction and management

Two review authors will independently extract data from eligible papers onto a bespoke data extraction form. The form will describe population tested, purpose/setting of testing, test(s) administered, details of person performing testing, and details of reference standard assessment. We will derive components of the 2x2 table and prevalence figures.

We will pilot the data extraction form on two papers and make any required changes. Following data extraction, the two review authors will compare their findings and discuss and resolve any disagreements, with recourse to a senior test accuracy review author as needed.

Where a test assigns a score and accuracy data are given for a variety of threshold scores, we will collect all of these data in the first instance. Primary analyses will be limited to performance at the standard threshold for that test (where a standard exists). Exploratory analyses will describe accuracy at other thresholds.

Assessment of methodological quality

Paired, independent raters blinded to each other's scores will assess risk of bias (internal validity) and generalisability (external validity) using the QUADAS-2 tool (www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/).

QUADAS-2 assessment covers issues relating to patient selection, index test, reference standard, and participant flow. Each domain is assessed for issues related to risk of bias; the first three domains are also assessed for generalisability concerns. Our group has considerable experience using the QUADAS-2 tool. For previous DTA reviews, we convened a group with expertise in test accuracy and dementia to tailor the QUADAS-2 approach to the field of dementia studies. Through this work we have operationalised scoring rules for item- and domain-level QUADAS-2 assessment and modified the generic QUADAS-2 question stems to cover issues pertinent to cognitive testing (Appendix 2) (Davis 2013).

QUADAS-2 results will be presented as graphical displays and as a narrative within the text. If data allow, we will perform a sensitivity analysis limiting to those papers with low concern for risk of bias across all the QUADAS domains.

Statistical analysis and data synthesis

Our primary analysis of interest is accuracy of the various remote assessments against the dichotomous outcome variable 'dementia/no dementia'. To explore this, we will apply the recommended Cochrane framework treating each test separately. For each test, where data allow, we will extract data to populate a standard 2x2 data table of binary test results (above and below threshold score) cross-classified against binary reference standard.

From this table we will calculate sensitivities and specificities, with 95% confidence intervals, at individual test level for thresholds of interest. Primary thresholds of interest will be based on the thresholds proposed in the original paper describing the study, unless there is a consensus agreed upon threshold that is used in practice and differs from the original threshold described. We will present these study results graphically in a forest plot of sensitivity and corresponding specificity. We will perform these first analyses with standard Review Manager 5 software (Review manager 5.3 2020).

Where there are more than two studies describing a test with the same threshold value, we will attempt quantitative meta-analysis. In the first instance we will create summary estimates of sensitivity and specificity using random-effects models and the bivariate approach. We will use bespoke software MetaDTA for this (Freeman 2019). Our approach to analysis assumes that the tests will use a common threshold to define a test positive case. If multiple thresholds are described for a single test, we will explore the potential for analysis using the summary receiver operating characteristic curve (SROC) approach. We recognise that there are new approaches to studies with different thresholds in the same study which make better use of the study-level data (Jones 2019). Regardless of approach used, we will create summary estimates with corresponding confidence intervals and range of maximum and minimum input sensitivity and specificity.

For assessment of remote versus in-person testing, if data are available as correlations or reliability or agreement measures, we will tabulate these and describe them in the narrative of the review but we will not attempt to create summary estimates. If scores are presented, we will describe scores and the proportion classified as having cognitive impairment for each test.

Investigations of heterogeneity

We will not quantify statistical heterogeneity for these DTA analyses. As a measure of the uncertainty around any summary estimate, we will present both 95% confidence and prediction intervals.

We will review forest plots of sensitivity and specificity looking for outliers.

Assuming that heterogeneity is present and that data allow, we will perform subgroup analyses to explore potential areas of heterogeneity that are common to DTA studies in the dementia field.

We will assess:

- populations tested, performing subgroup analyses if specific disease groups feature in the included papers (e.g. traumatic brain injury, stroke), or if specific healthcare settings feature (e.g. secondary care clinics, where prevalence of disease will differ to other settings such as primary care);
- technical features of the testing strategy (platform used for delivering the test; language of testing; person performing testing);
- clinical criteria used to reach dementia diagnosis (e.g. ICD-10; DSM-IV) and the methodology used to reach dementia diagnosis (e.g. individual assessment; group (consensus) assessment).

Sensitivity analyses

Where appropriate (i.e. if not already explored in our analyses of heterogeneity) and as data allow, we will explore the effect of methodological aspects of the included studies. In the first instance we will run a sensitivity analysis limited to studies at low risk of bias.

Assessment of reporting bias

We will not perform a quantitative assessment of reporting bias. We recognise that debate remains around the most robust approach to assessment of reporting bias in DTA (Wilson 2015), and there is uncertainty on how to apply standard approaches such as funnel plots (Annefloor van Enst 2014).

ACKNOWLEDGEMENTS

We would like to thank peer reviewers Andrew Larner and Dimity Pond and consumer reviewer Cathie Hofstetter for their comments and feedback.

REFERENCES

Additional references

Annefloor van Enst 2014

Annefloor van Enst W, Ochodo E, Scholten RJPM, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC Medical Research Methodology* 2014;**14**:70.

APA 2013

American Psychiatric Association, DSM-5 Task Force. Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th edition. Vol. **xliv**. Washington, DC: American Psychiatric Association, 2013.

Barth 2018

Barth J, Nickel F. Diagnosis of cognitive decline and dementia in rural areas - a scoping review. *International Journal of Geriatric Psychiatry* 2018;**33**(3):459-74.

Beishon 2019

Beishon LC, Batterham AP, Quinn TJ, Nelson CP, Panerai RB, Robinson T, et al. Addenbrooke's Cognitive Examination III (ACE-III) and mini-ACE for the detection of dementia and mild cognitive impairment. *Cochrane Database of Systematic Reviews* 2019, Issue 12. Art. No: CD013282. [DOI: 10.1002/14651858.CD013282.pub2]

Brandt 1988

Brandt J, Spencer M, Folstein M. The Telephone Interview for Cognitive Status. *Neuropsychiatry, Neuropsychology and Behavioral Neurology* 1988;**1**(17):111-7.

Covidence 2020 [Computer program]

Veritas Health Innovation Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation. www.covidence.org.

Creavin 2016

Creavin ST, Wisniewski S, Noel-Storr AH, Trevelyan CM, Hampton T, Rayment D, et al. Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *Cochrane Database of Systematic Reviews* 2016, Issue 1. Art. No: CD011145. [DOI: 10.1002/14651858.CD011145.pub2]

Davis 2013

Davis DHJ, Creavin ST, Noel-Storr A, Quinn TJ, Smailagic N, Hyde C, 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 of Systematic Reviews* 2013, Issue 3. Art. No: CD010460. [DOI: 10.1002/14651858.CD010460]

Davis 2015

Davis DHJ, Creavin ST, Yip JLY, Noel-Storr AH, Brayne C, Cullum S. Montreal Cognitive Assessment for the diagnosis of Alzheimer's disease and other dementias. *Cochrane Database of Systematic Reviews* 2015, Issue 10. Art. No: CD010775. [DOI: 10.1002/14651858.CD010775.pub2]

Deeks 2001

Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;**323**:157-62.

Elliott 2020

Elliott E, Green C, Llewellyn DJ, Quinn TJ. Accuracy of telephone-based cognitive screening tests: systematic review and meta-analysis. Current Alzheimer's Research (in press).

Freeman 2019

Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Medical Research Methodology* 2019;**19**(1):81.

Harrison 2015

Harrison JK, Fearon P, Noel-Storr AH, McShane R, Stott DJ, Quinn TJ. Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a secondary care setting. *Cochrane Database of Systematic Reviews* 2015, Issue 3. Art. No: CD010772. [DOI: 10.1002/14651858.CD010772.pub2.]

Harrison 2016

Harrison K, Noel-Storr AH, Demeyere N, Reynish EL, Quinn TJ. Outcomes measures in a decade of dementia and mild cognitive impairment trials. *Alzheimer's Research & Therapy* 2016;**8**(1):48.

Hendry 2016

Hendry K, Quinn TJ, Evans J, Scortichini V, Miller H, Burns J, et al. Evaluation of delirium screening tools in geriatric medical inpatients: a diagnostic test accuracy study. *Age and Ageing* 2016;**45**:832-7.

Hendry 2019

Hendry K, Green C, McShane R, Noel-Storr AH, Stott DJ, Anwer S, et al. AD-8 for detection of dementia across a variety of healthcare settings. *Cochrane Database of Systematic Reviews* 2019, Issue 3. Art. No: CD011121. [DOI: DOI: 10.1002/14651858.CD011121.pub2]

Jones 2019

Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Statistics in Medicine* 2019;**38**:4789-803.

Lees 2012

Lees R, Fearon P, Harrison JK, Broomfield NM, Quinn TJ. Cognitive and mood assessment in stroke research: focused review of contemporary studies. *Stroke* 2012;**43**(6):1678-80.

Lin 2013

Lin JS, O'Connor E, Rossom RC, Perdue LA, Eckstrom E. Screening for cognitive impairment in older adults: a systematic

review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine* 2013;**159**(9):601-12.

Noel-Storr 2012

Noel-Storr AH, McCleery JM, Richard E, Ritchie CW, Flicker L, Cullum SJ, et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. *Neurology* 2012;**83**(4):364-73.

Owen 2018

Owen RK, Cooper NJ, Quinn TJ, Lees R, Sutton AJ. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. *Journal of Clinical Epidemiology* 2018;**99**:64-74.

Review manager 5.3 2020 [Computer program]

The Cochrane Collaboration Review manager 5.3. The Cochrane Collaboration, 2020. revman.cochrane.org.

Ritchie 2015

Ritchie CW, Terrera GM, Quinn TJ. Dementia trials and dementia tribulations: methodological and analytical challenges in dementia research. *Alzheimer's Research & Therapy* 2015;**7**(1):31.

Robinson 2015

Robinson L, Tang E, Taylor JP. Dementia: timely diagnosis and early intervention. *BMJ* 2015;**350**:h3029.

Takwoingi 2018

Takwoingi Y, Quinn TJ. Review of Diagnostic Test Accuracy (DTA) studies in older people. *Age and Ageing* 2018;**47**(3):349-55.

Whiting 2008

Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic reviews of test accuracy should search a range of databases to identify primary studies. *Journal of Clinical Epidemiology* 2008;**41**(4):357-64.

WHO 2010

World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD). Vol. **2**. Geneva: World Health Organization, 2010.

Wilson 2015

Wilson C, Kerr D, Noel-Storr A, Quinn TJ. Associations with publication and assessing publication bias in dementia diagnostic test accuracy studies. *International Journal of Geriatric Psychiatry* 2015;**30**(12):1250-6.

APPENDICES

Appendix 1. MEDLINE search strategy

1 exp DEMENTIA/

2 major cognitive disorder.ti,ab.

3 alzheimer*.ti,ab.

4 dement*.ti,ab.

5 ((lewy adj2 bod*) or LBD or DLB).ti,ab.

6 (FTLD or frontotemp*).ti,ab.

7 or/1-6

8 exp Neuropsychological Tests/

9 exp Cognition Disorders/di [Diagnosis]

10 ((cognit* or memor* or neuropsychological*) adj3 (assess* or test* or task* or performance* or decline* or function*)).ti,ab.

11 MoCA.ti,ab.

12 MMSE.ti,ab.

13 "Mini-mental State Examination".ti,ab.

14 "Brief Screen for Cognition Impairment".ti,ab.

15 "Memory and Ageing Telephone Screen".ti,ab.

16 "Telephone Cognitive Assessment Battery".ti,ab.

17 "Short Portable Mental Status Questionnaire".ti,ab.

18 "Telephone Modified Mini- Mental state exam".ti,ab.

19 "Telephone administered Minnesota Cognitive Acuity Screen".ti,ab.

20 "Blessed Telephone Information Memory Concentration Test".ti,ab.

21 "Structured telephone interview for dementia assessment".ti,ab.

22 TICS.ti,ab.

23 TICSm.ti,ab.

24 TICS-M.ti,ab.

25 sMMSE.ti,ab.

26 "Telephone Interview for Cognitive Status".ti,ab.

27 or/8-26

28 7 and 27

29 exp Internet/

30 Smartphone/

31 exp Telecommunications/

32 camera*.ti,ab.

33 phone*.ti,ab.

34 Smartphone.ti,ab.

35 teleconferenc*.ti,ab.

36 telephone*.ti,ab.

37 telepsychiatry.ti,ab.

38 telemedicine*.ti,ab.

39 video*.ti,ab.

40 webcam*.ti,ab.

41 (remote* adj (test* or diagnos* or consult* or deliver*)).ti,ab.

42 "mobile tablet".ti,ab.

43 or/29-42

44 28 and 43

Appendix 2. QUADAS-2 anchoring statements

We provide some core anchoring statements for quality assessment of diagnostic test accuracy reviews of neuropsychological tests in dementia. These statements are designed for use with the QUADAS-2 tool and were derived during a two-day, multidisciplinary focus group in 2010. If a QUADAS-2 signalling question for a specific domain is answered 'yes', the risk of bias can be judged to be 'low'. If a question is answered 'no', this indicates a risk of potential bias. The focus group was tasked with judging the extent of the bias for each domain. During this process, it became clear that certain issues were key to assessing quality, whilst others were important to record but were less important for assessing overall quality. To assist, we describe a 'weighting' system. When an item is weighted 'high risk', that section of the QUADAS-2 results table is judged to have a high potential for bias if a signalling question is answered 'no'. For example, in dementia diagnostic test accuracy studies, ensuring that clinicians performing dementia assessment are blinded to results of the index test is fundamental. If this blinding was not present, the item on the reference standard should be scored 'high risk of bias', regardless of

the other contributory elements. When an item is weighted 'low risk', it is judged to have a low potential for bias if a signalling question for that section of the QUADAS-2 results table is answered 'no'. Overall bias will be judged on whether other signalling questions (with a high risk of bias) for the same domain are also answered 'no'. In assessing individual items, a score of 'unclear' should be given only if there is genuine uncertainty. In these situations, the review authors will contact the relevant study teams for additional information.

Anchoring statements to assist with 'Risk of bias' assessment

Domain 1: Patient selection

Risk of bias: could the selection of patients have introduced bias? (high/low/unclear)

Was a consecutive or random sample of patients enrolled? When sampling is used, the methods least likely to cause bias are consecutive sampling and random sampling, which should be stated or described, or both. Non-random sampling or sampling based on volunteers is more likely to be at high risk of bias. Weighting: high risk of bias

Was a case-control design avoided? Case-control study designs have a high risk of bias, but are sometimes the only studies available, especially if the index test is expensive or invasive, or both. Nested case-control designs (systematically selected from a defined population cohort) are less prone to bias, but they will still narrow the spectrum of patients that receive the index test. Study designs (both cohort and case-control) that may also increase bias are those designs in which the study team deliberately increases or decreases the proportion of participants with the target condition, for example a population study may be enriched with extra dementia participants from a secondary care setting. Weighting: high risk of bias

Did the study avoid inappropriate exclusions? The study will be automatically graded as unclear if exclusions are not detailed (pending contact with study authors). When exclusions are detailed, we will grade the study as 'low risk' if we feel that the exclusions are appropriate. Certain exclusions common to many studies of dementia are medical instability, terminal disease, alcohol/substance misuse, concomitant psychiatric diagnosis, and other neurodegenerative condition. However, if 'difficult to diagnose' groups are excluded, this may introduce bias, so exclusion criteria must be justified. For a community sample, we would expect relatively few exclusions. We will label post hoc exclusions as 'high risk' of bias. Weighting: high risk of bias

Applicability: are there concerns that the included patients do not match the review question? (high/low/unclear)

The included patients should match the intended population as described in the review question. If not already specified in the review inclusion criteria, the setting will be particularly important – the review authors should consider population in terms of symptoms, pretesting, and potential disease prevalence. We will classify studies that use very selected participants or subgroups as low applicability, unless they are intended to represent a defined target population, for example people with memory problems referred to a specialist and investigated by lumbar puncture.

Domain 2: Index test

Risk of bias: could the conduct or interpretation of the index test have introduced bias? (high/low/unclear)

Were the index test results interpreted without knowledge of the reference standard? Terms such as 'blinded' or 'independently and without knowledge of' are sufficient; full details of the blinding procedure are not required. This item may be scored as 'low risk' if it is explicitly described, or if there is a clear temporal pattern to the order of testing that precludes the need for formal blinding (e.g. all (neuropsychological test) assessments were performed before the dementia assessment). As most neuropsychological tests are administered by a third party, knowledge of dementia diagnosis may influence their ratings; tests that are self-administered, for example by using a computerised version, may have less risk of bias. Weighting: high risk

Were the index test thresholds prespecified? For neuropsychological scales, there is usually a threshold above which participants are classified as 'test positive'; this may be referred to as threshold, clinical cut-off, or dichotomisation point. Different thresholds are used in different populations. A study is classified as at higher risk of bias if the authors define the optimal cut-off post hoc based on their own study data. Certain papers may use an alternative methodology for analysis that does not use thresholds; these papers should be classified as not applicable. Weighting: low risk

Were sufficient data on (neuropsychological test) application given for the test to be repeated in an independent study? Particular points of interest include method of administration (e.g. self-completed questionnaire versus direct-questioning interview), nature of informant, and language of assessment. If a novel form of the index test is used, for example a translated questionnaire, details of the scale should be included and a reference given to an appropriate descriptive text, and evidence of validation should be provided. Weighting: low risk

Applicability: are there concerns that the index test, its conduct, or its interpretation may differ from the review question? (high/low/unclear)

Variations in the length, structure, language, and/or administration of the index test may all affect applicability if they differ from those specified in the review question.

Domain 3: Reference standard

Risk of bias: could the reference standard, its conduct, or its interpretation have introduced bias? (high/low/unclear)

Is the reference standard likely to correctly classify the target condition? Commonly used international criteria that can assist with clinical diagnosis of dementia include those detailed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and the International Classification of Diseases (ICD-10). Criteria specific to dementia subtypes include but are not limited to the National Institute of Neurological and Communicative Diseases and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria for Alzheimer's dementia; McKeith criteria for Lewy body dementia; Lund criteria for frontotemporal dementia; and National Institute of Neurological Disorders and Stroke and Association Internationale pour la Recherche et l'Enseignement en Neurosciences (NINDS-AIREN) criteria for vascular dementia. When the criteria used for assessment are unfamiliar to the review authors and the Cochrane Dementia and Cognitive Improvement Group, this item should be classified as 'high risk of bias'. Weighting: high risk

Were the reference standard results interpreted without knowledge of the results of the index test? Terms such as 'blinded' or 'independent' are sufficient; full details of the blinding procedure are not required. This may be scored as 'low risk' if explicitly described, or if a clear temporal pattern to the order of testing is evident (e.g. all dementia assessments performed before (neuropsychological test) testing). Informant rating scales and direct cognitive tests present certain problems. It is accepted that informant interview and cognitive testing are usual components of clinical assessment for dementia; however, specific use of the scale under review in the clinical dementia assessment should be scored as high risk of bias. Weighting: high risk

Was sufficient information on the method of dementia assessment given for the assessment to be repeated in an independent study? Particular points of interest for dementia assessment include the training/expertise of the assessor; whether additional information (e.g. neuroimaging; other neuropsychological test results) was available to inform the diagnosis; and whether this was available for all participants. Weighting: variable risk, but high risk if method of dementia assessment not described

Applicability: are there concerns that the target condition as defined by the reference standard does not match the review question? (high/low/unclear)

There exists the possibility that some methods of dementia assessment, although valid, may diagnose a smaller or larger proportion of participants with disease than in usual clinical practice. In these instances, the item should be rated 'poor applicability'.

Domain 4: Patient flow and timing (N.B. refer to, or construct, a flow diagram)

Risk of bias: could the patient flow have introduced bias? (high/low/unclear)

Was there an appropriate interval between the index test and the reference standard? For a cross-sectional study design, the potential exists for the participant to change between assessments; however, dementia is a slowly progressive disease that is not reversible. The ideal scenario would be a same-day assessment, but longer periods of time (e.g. several weeks or months) are unlikely to lead to a high risk of bias. For delayed-verification studies, the index and reference tests are necessarily separated in time, given the nature of the condition. Weighting: low risk

Did all participants receive the same reference standard? In some scenarios, participants who score 'test positive' on the index test have a more detailed assessment for the target condition. When dementia assessment (or the reference standard) differs between participants, this should be classified as high risk of bias. Weighting: high risk

Were all participants included in the final analysis? Attrition will vary with study design. Delayed-verification studies will have higher attrition than cross-sectional studies because of mortality, and this is likely to be greater in participants with the target condition. Dropouts (and missing data) should be accounted for. Attrition that is higher than expected (compared with other similar studies) should be treated as high risk of bias. We have defined a cut-off of greater than 20% attrition as being high risk, but this will be highly dependent on the length of follow-up in individual studies. Weighting: high risk

HISTORY

Protocol first published: Issue 9, 2020

CONTRIBUTIONS OF AUTHORS

Terry Quinn and Jenny McLeery conceived the idea. All authors contributed to writing the protocol and responding to peer review comments. The Cochrane Dementia and Cognitive Improvement Group Information Specialists designed the search.

DECLARATIONS OF INTEREST

Terry J Quinn: none known
Emma Elliott: none known
Tuuli M Hietamies: none known
Gabriel Martínez: none known
Zoë Tieges: none known

Riona Mc Ardle: none known

SOURCES OF SUPPORT

Internal sources

- No sources of support supplied

External sources

- NIHR, UK

This protocol was supported by the National Institute for Health Research (NIHR), via Cochrane Infrastructure funding to the Cochrane Dementia and Cognitive Improvement Group. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Systematic Reviews Programme, NIHR, National Health Service, or the Department of Health