Cochrane Library

Cochrane Database of Systematic Reviews

Clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people (Review)

Creavin ST, Noel-Storr AH, Langdon RJ, Richard E, Creavin AL, Cullum S, Purdy S, Ben-Shlomo Y

Creavin ST, Noel-Storr AH, Langdon RJ, Richard E, Creavin AL, Cullum S, Purdy S, Ben-Shlomo Y.
Clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people.
Cochrane Database of Systematic Reviews 2022, Issue 6. Art. No.: CD012558.
DOI: 10.1002/14651858.CD012558.pub2.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Diagnostic Test Accuracy Review]

Clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people

Samuel T Creavin[1], Anna H Noel-Storr[2], Ryan J Langdon[3], Edo Richard[4], Alexandra L Creavin[1], Sarah Cullum[5], Sarah Purdy[1], Yoav Ben-Shlomo[1]

[1]Population Health Sciences, University of Bristol, Bristol, UK. [2]Radcliffe Department of Medicine, University of Oxford , Oxford, UK. [3]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [4]Department of Neurology, Donders Institute for Brain, Behaviour and Cognition, Radboud University Nijmegen Medical Center, Nijmegen, Netherlands. [5]Department of Psychological Medicine, University of Auckland, Auckland, New Zealand

Contact: Samuel T Creavin, sam.creavin@bristol.ac.uk.

Editorial group: Cochrane Dementia and Cognitive Improvement Group.
Publication status and date: New, published in Issue 6, 2022.

Citation: Creavin ST, Noel-Storr AH, Langdon RJ, Richard E, Creavin AL, Cullum S, Purdy S, Ben-Shlomo Y. Clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people. *Cochrane Database of Systematic Reviews* 2022, Issue 6. Art. No.: CD012558. DOI: 10.1002/14651858.CD012558.pub2.

ABSTRACT

Background

In primary care, general practitioners (GPs) unavoidably reach a clinical judgement about a patient as part of their encounter with patients, and so clinical judgement can be an important part of the diagnostic evaluation. Typically clinical decision making about what to do next for a patient incorporates clinical judgement about the diagnosis with severity of symptoms and patient factors, such as their ideas and expectations for treatment. When evaluating patients for dementia, many GPs report using their own judgement to evaluate cognition, using information that is immediately available at the point of care, to decide whether someone has or does not have dementia, rather than more formal tests.

Objectives

To determine the diagnostic accuracy of GPs' clinical judgement for diagnosing cognitive impairment and dementia in symptomatic people presenting to primary care. To investigate the heterogeneity of test accuracy in the included studies.

Search methods

We searched MEDLINE (Ovid SP), Embase (Ovid SP), PsycINFO (Ovid SP), Web of Science Core Collection (ISI Web of Science), and LILACs (BIREME) on 16 September 2021.

Selection criteria

We selected cross-sectional and cohort studies from primary care where clinical judgement was determined by a GP either prospectively (after consulting with a patient who has presented to a specific encounter with the doctor) or retrospectively (based on knowledge of the patient and review of the medical notes, but not relating to a specific encounter with the patient). The target conditions were dementia and cognitive impairment (mild cognitive impairment and dementia) and we included studies with any appropriate reference standard such as the Diagnostic and Statistical Manual of Mental Disorders (DSM), International Classification of Diseases (ICD), aetiological definitions, or expert clinical diagnosis.

Data collection and analysis

Two review authors screened titles and abstracts for relevant articles and extracted data separately with differences resolved by consensus discussion. We used QUADAS-2 to evaluate the risk of bias and concerns about applicability in each study using anchoring statements. We performed meta-analysis using the bivariate method.

Main results

We identified 18,202 potentially relevant articles, of which 12,427 remained after de-duplication. We assessed 57 full-text articles and extracted data on 11 studies (17 papers), of which 10 studies had quantitative data. We included eight studies in the meta-analysis for the target condition dementia and four studies for the target condition cognitive impairment. Most studies were at low risk of bias as assessed with the QUADAS-2 tool, except for the flow and timing domain where four studies were at high risk of bias, and the reference standard domain where two studies were at high risk of bias. Most studies had low concern about applicability to the review question in all QUADAS-2 domains.

Average age ranged from 73 years to 83 years (weighted average 77 years). The percentage of female participants in studies ranged from 47% to 100%. The percentage of people with a final diagnosis of dementia was between 2% and 56% across studies (a weighted average of 21%). For the target condition dementia, in individual studies sensitivity ranged from 34% to 91% and specificity ranged from 58% to 99%. In the meta-analysis for dementia as the target condition, in eight studies in which a total of 826 of 2790 participants had dementia, the summary diagnostic accuracy of clinical judgement of general practitioners was sensitivity 58% (95% confidence interval (CI) 43% to 72%), specificity 89% (95% CI 79% to 95%), positive likelihood ratio 5.3 (95% CI 2.4 to 8.2), and negative likelihood ratio 0.47 (95% CI 0.33 to 0.61).

For the target condition cognitive impairment, in individual studies sensitivity ranged from 58% to 97% and specificity ranged from 40% to 88%. The summary diagnostic accuracy of clinical judgement of general practitioners in four studies in which a total of 594 of 1497 participants had cognitive impairment was sensitivity 84% (95% CI 60% to 95%), specificity 73% (95% CI 50% to 88%), positive likelihood ratio 3.1 (95% CI 1.4 to 4.7), and negative likelihood ratio 0.23 (95% CI 0.06 to 0.40).

It was impossible to draw firm conclusions in the analysis of heterogeneity because there were small numbers of studies. For specificity we found the data were compatible with studies that used ICD-10, or applied retrospective judgement, had higher reported specificity compared to studies with DSM definitions or using prospective judgement. In contrast for sensitivity, we found studies that used a prospective index test may have had higher sensitivity than studies that used a retrospective index test.

Authors' conclusions

Clinical judgement of GPs is more specific than sensitive for the diagnosis of dementia. It would be necessary to use additional tests to confirm the diagnosis for either target condition, or to confirm the absence of the target conditions, but clinical judgement may inform the choice of further testing. Many people who a GP judges as having dementia will have the condition. People with false negative diagnoses are likely to have less severe disease and some could be identified by using more formal testing in people who GPs judge as not having dementia. Some false positives may require similar practical support to those with dementia, but some - such as some people with depression - may suffer delayed intervention for an alternative treatable pathology.

PLAIN LANGUAGE SUMMARY

Why is improving dementia diagnosis important?

Dementia refers to a group of brain conditions that lead to progressive problems with memory, working-things-out, or functioning in everyday life. Doctors use a variety of tests to diagnose dementia. People have often reported that it can take a long time to get a diagnosis of dementia from initially presenting to a healthcare provider with symptoms suggestive of dementia.

Cognitive impairment is a broader term that includes people whose brain is not functioning as well as expected given their age, but they do not have dementia, as well as people with dementia. Some people with cognitive impairment who do not have dementia may have a condition called mild cognitive impairment (MCI). Some people with MCI (but not all) will develop dementia over time.

What is the aim of this review?

The review authors aimed to investigate the diagnostic accuracy of clinical judgement of general practitioners (GPs) for diagnosing dementia, and cognitive impairment, in primary care.

What was studied in the review?

The authors included extracted data from 11 studies, including 10 with complete data on diagnostic accuracy. The authors included eight studies in the statistical summary with a total of 2790 people, of whom 826 (30%) had dementia. The authors included four studies that investigated cognitive impairment as the condition to diagnose, with a total of 1497 people of whom 594 had cognitive impairment (40%).

What are the main results of the review?

The results of the review indicate that in theory, if GPs used their clinical judgement in practice for dementia, they would correctly identify 58% of the people who have dementia as having the condition (sensitivity) and 89% of the people who do not have dementia as being free of the condition (specificity).

The results of the review indicate that in theory, if GPs used their clinical judgement in practice for cognitive impairment, they would correctly identify 84% of the people who have cognitive impairment as having the condition (sensitivity) and 73% of the people who do not have cognitive impairment as being free of the condition (specificity).

How reliable are the results of the studies in this review?

In this review there were some technical problems with the design of the original studies, and there were differences between studies that made it difficult to compare them to each other. This means that it is difficult to be certain how applicable these findings are in clinical practice.

Who do the results of this review apply to?

Researchers who conducted the studies in the review carried out their investigations mostly in Europe, with one study in the USA and one study in Australia. All studies included people attending their GP. Average age ranged from 73 years to 83 years (weighted average 77 years). The percentage of female participants in studies ranged from 47% to 100%. The percentage of people with a final diagnosis of dementia was between 2% and 56% across studies (a weighted average of 21%). If applying these findings in settings with fewer number of people with dementia then the accuracy of the test may be different.

What are the implications of this review?

If these studies are indeed representative of GPs practice, then if GPs used their clinical judgement alone to diagnose dementia then this might mean that some people with dementia are incorrectly 'missed', and it is important to do further tests to confirm that the person does not have a problem. However, if a GP thinks someone has dementia there is a good chance that the diagnosis is correct and the test to confirm dementia might be different and potentially less time consuming and burdensome. The studies included in this review suggest that clinical judgement could be a useful test to determine what to do next.

How up-to-date is this review?

The review authors searched for and used studies published up to 16 September 2021.

SUMMARY OF FINDINGS

Summary of findings 1. Summary of findings table

Title: clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people

Population: adults without an existing diagnosis of dementia, registered with general practices

Setting: primary care, defined as first-contact health care provided by a non-specialist clinician in a continuing-care office setting

Index test: clinical judgement of a general practitioner (GP), defined as being unaided by any additional test, investigation or inquiry beyond that which is immediately available to the clinician

Reference standard: for dementia, diagnosis of dementia using any recognised classification systems. For cognitive impairment, diagnosis of mild cognitive impairment (MCI) or other cognitive impairment (including dementia) using any recognised classification system

Studies: cross-sectional studies and cohort studies, case-control studies were excluded

Limitations: because of the small numbers there was uncertainty in the findings as indicated by the wide confidence intervals (CI)

Test	Summary accuracy (95% CI)	Number of participants	Prevalence of target condition	Quality, implications, comments
Clinical judgement of general practitioners for the diagnosis of dementia (8 studies)	Sensitivity 58% (95% CI 43% to 72%) specificity 89% (95% CI 79% to 95%) positive likelihood ratio 5.3 (95% CI 2.4 to 8.2) negative likelihood ratio 0.47 (95% CI 0.33 to 0.61)	826 out of 2790 had dementia	30% (95% CI 28% to 31%)	Studies were generally at low risk of bias with the exception of the flow and timing domain, which relates to the possibility of selectively confirming dementia status in people depending on the basis of the results of GP judgement; this increases the possibility of misclassifying mild cases of dementia as true negative, when in fact they are false negative. There were low concerns about applicability. Judgement is likely to be more accurate at confirming diagnosis than excluding dementia
Clinical judgement of general practitioners for the diagnosis of cognitive impairment (4 studies)	Sensitivity 84% (95% CI 60% to 95%) specificity 73% (95% CI 50% to 88%) positive likelihood ratio 3.1 (95% CI 1.4 to 4.7) negative likelihood ratio 0.23 (95% CI 0.06 to 0.40)	594 out of 1497 had cognitive impairment	40% (95% CI 37% to 42%)	Studies were generally at low risk of bias with the exception of the flow and timing domain. There were low concerns about applicability. Judgement is likely to be more accurate at excluding cognitive impairment than confirming diagnosis

CAUTION: the results on this table should not be interpreted in isolation from the results of the individual included studies contributing to each summary test accuracy measure. These are reported in the main body of the text of the review

BACKGROUND

Doctors use a variety of processes to reach a diagnosis, including non-analytical reasoning processes such as pattern recognition, to rapidly generate diagnostic hypotheses (Elstein 2009; Norman 2007). Clinical judgement of general practitioners has recently been shown to be of value in predicting functional decline in community dwelling older adults (van Blijswijk 2018), and in evaluating chest pain (Harskamp 2019). Some people with dementia unfortunately have sufficiently advanced disease at the point of diagnosis that additional tests or formal evaluation may be unnecessary and burdensome. General practitioners (GPs) often report using their clinical judgement, rather than a formal test, to determine whether someone has dementia (O'Connor 1993; Pentzek 2009). Some authors have suggested that clinical judgement of general practitioners might improve calibration of models for future dementia risk prediction (Ford 2018; Pentzek 2019).

Target condition being diagnosed

Dementia is a clinical syndrome of cognitive impairment that develops gradually and causes a decline in functioning. Dementia is increasingly common with age, affecting less than 5% of the population aged less than 75 years and 17% of those aged over 89 years (Matthews 2013). Dementia may result from a variety of pathologies, but in the elderly population in the community some investigators believe that subtype definitions based on disease aetiology are of less relevance, as most old people with dementia have mixed pathology at autopsy (Brayne 2012; Kawas 2015; Neuropathology 2001; Savva 2009).

Cognitive impairment includes dementia and mild cognitive impairment (MCI) (Gauthier 2006). MCI is a syndrome of cognitive impairment that is greater than expected when accounting for a person's age and educational attainment, but that does not significantly interfere with capacity for independence in everyday activities of daily living. MCI affects between 3% and 20% of adults aged over 65 years (Gauthier 2006), and the prognosis in general practice is variable: approximately 25% of people develop dementia within 3 years but around 40% revert to normal (Kaduszkiewicz 2014).

Index test(s)

The index test was a clinical judgement of a primary care physician after a clinical assessment, unaided by formal (even brief) cognitive tests. This was operationalised as a single index test (clinical judgement) with two Target conditions.

The diagnostic process can be scrutinised as a probabilistic, analytic process with a clear diagnostic category as the intended outcome. In general practice, diagnostic labels may function primarily to guide the management of the patient, to treat, to investigate, or to exclude serious disease (Jones 2010). In clinical practice, decisions are typically dichotomous (Croft 2015), and reasoning is often non-analytic (Brush 2017; Kahneman 2011), based on intuition which is derived from recognition (Brush 2017). Clinical judgement, also known as instinct, gut feeling, or gestalt (Lambe 2016; Lehman 2015), is the holistic judgement about the diagnostic category. Understood in this framework, diagnosis is typically a System one cognitive process. Contrasted to System two (conscious, deliberate, explicit, effortful), System one is mostly unconscious, involuntary, implicit, and low-effort, but is prone to a range of biases, such as neglecting ambiguity, ignoring absent evidence, overestimating low probabilities, and using heuristics (Kahneman 2011). Authors have described GPs as using intuition (Barraclough 2006; Woolley 2013), pattern recognition (Heneghan 2009), and scripts (Charlin 2000), amongst other strategies (Heneghan 2009), to reach a diagnosis.

GPs report lack of time as a barrier to diagnosing dementia (Koch 2010) and report often relying on personal observations to make the diagnosis (O'Connor 1993) whereas pen-and-paper tests are used by a minority of people (Pond 2013).

Clinical pathway

Prior tests

Most commonly no tests would be performed before a GP consultation regarding possible dementia because the barriers to consultation with a primary care physician are relatively low in most health economies. Some people may consult with a healthcare professional because they have subjective cognitive impairment and are concerned about the onset of possible dementia, and typically the first encounter would be with a primary care provider (commonly a GP), but in some health economies the first consultation may be with a specialist clinician. Other people may not experience subjective cognitive problems (Waldorff 2012), but a close contact (e.g. a carer) or professional who is concerned about possible dementia may encourage (or take) the person who they are concerned about to attend a consultation with a clinician. A further possibility is that a GP may form an impression of possible cognitive impairment during a consultation with a patient about a (potentially) unrelated matter.

Some people may consult with their GP about the possibility of dementia after performing a self-administered cognitive test such as Test-Your-Memory (Brown 2009), although the authors of this test are explicit that it should be completed under medical supervision, or other cognitive testing. Alternatively some people might have been asked to see their GP as a consequence of undergoing brief cognitive testing conducted by another health professional (e.g. a district nurse or hospital doctor), or as part of a research project. The availability of test results to the consulting clinician is likely to be variable.

In this review we only considered clinical judgement by a primary care physician (GP) in someone who has symptoms. Either the patient themselves or someone else, including a health professional (including the consulting GP), must have been concerned about possible cognitive impairment. Recent policy in the USA and UK has encouraged case-finding for dementia in people who do not have symptoms (Burns 2013; Rasmussen 2013; Rasmussen 2014). This remains controversial (Brayne 2007; Fox 2013; Iliffe 2014; Le 2013) and we did not plan to include studies specifically addressing case-finding in this review.

Role of index tests

It is rare that any single component of a diagnostic evaluation would be diagnostic for a condition by itself. GPs who are evaluating people for possible cognitive impairment or dementia may request further assessment such as brief cognitive tests (such as General Practitioner Assessment of Cognition (GPCOG) (Brodaty 2002), or Montreal Cognitive Assessment (MoCA) (Nasreddine 2005), and investigations such as biochemical analysis and neuroimaging.

Depending on the clinical scenario there may also be further workup to exclude alternative diagnoses such as depression. GPs' clinical judgement is often important in determining what further assessment they request: when the GP feels comfortable to exclude cognitive impairment or dementia without further assessment the patient will usually undergo no further tests, whereas the GP may arrange further evaluation if they feel uncertain. It would be unusual for a GP to rule-in dementia without further assessment, but this may occur when the patient is frail, affected by multiple comorbidities, and perhaps resident in a nursing home, where the prior probability of dementia (or prevalence) may be as high as 60% (Magaziner 2000), and when the management may be primarily palliative. In some situations GPs may use a test of time (Almond 2009) to help increase the specificity of a diagnosis, especially when the condition may fluctuate, and a GP may therefore form a 'working diagnosis', which is reviewed over a period of time, before deciding on a formal recorded diagnosis.

The doctor's clinical judgement will typically determine the extent of the additional work-up offered. If a GP thinks someone is highly likely to have dementia, then the GP might only request a brief 'rule-in' test together with blood tests to exclude other causes such as hypothyroidism or infection, or (rarely) no additional tests at all; this scenario is less applicable to people who GPs think have cognitive impairment rather than frank dementia and mostly applies to people with advanced disease. If a GP thinks a person is highly unlikely to have any problems, then the GP might offer a brief 'rule-out' test, or none at all. A GP is likely to refer to a specialist if there is residual uncertainty after initial evaluation.

Alternative test(s)

Because in most health systems the barrier to consultation with a GP is low, and a GP necessarily and unavoidably forms a clinical judgement during an encounter with a patient, there are not alternative tests that would generally be applied to all consulters. As described above, it would be unusual (though not impossible) that the index test would be used in isolation, and often, though not always, additional tests would be performed following an initial assessment.

Alternatives to the index test would include a more detailed subsequent evaluation following initial GP judgement, perhaps done by a specialist, and might include aspects of clinical history, examination, cognitive testing, biochemical and haematological analysis, and neuroimaging.

Rationale

A systematic review published in 2012 found that the judgement of GPs was highly specific for diagnosing dementia at all stages of severity, but only moderately sensitive (van den Dungen 2012). A second review addressing a similar question used a more restricted search strategy (Mitchell 2011). The approach of both reviews was sound, but the inclusion criteria allowed for a broad definition of 'clinical judgement' that is not immediately applicable to clinical practice, because the review authors included studies that defined 'clinical judgement' as a documented diagnosis in the medical records, and this may not accurately reflect the actual clinical opinion (Russell 2013). Additionally the search strategy of both reviews was relatively restricted in particular regarding terms relating to dementia, cognitive impairment, and diagnostic accuracy.

Understanding the diagnostic accuracy of clinical judgement is important to help manage the clinical scenario of cognitive problems. Understanding the limitations of a GP judgement of dementia regarding sensitivity would help inform GPs on the importance of doing further tests to confirm normal cognition in someone who they believe does not have dementia, and therefore avoid missing cases, reducing uncertainty and distress from unexplained symptoms. In contrast, knowing the limitations of a GP diagnosis of dementia regarding specificity would help to inform the importance and value of doing further tests to confirm a diagnosis of dementia in someone who a GP suspects has dementia, in order to reduce false positives (who potentially have an alternative cause for their symptoms, such as depression). Understanding the accuracy of clinical judgement for cognitive impairment provides insights into the ability of GPs to identify people with this more broadly defined condition, with likely less severe symptoms than dementia.

OBJECTIVES

To determine the accuracy of general practitioners' (GPs) clinical judgement for diagnosing cognitive impairment and dementia in symptomatic people presenting to primary care. There is no comparator index test.

Secondary objectives

To investigate whether there was heterogeneity of test accuracy in the included studies for four aspects of study design: the reference standard; whether GP clinical judgement was prospective or retrospective; whether GPs had access to the medical records; and the risk of bias in the flow and timing Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) domain (see Investigations of heterogeneity for details).

METHODS

Criteria for considering studies for this review

Types of studies

We included cross-sectional studies and cohort studies. We included diagnostic accuracy studies of clinical judgement of general practitioners (GPs) that were 'nested' within epidemiological community-based dementia prevalence studies. Cross-sectional studies are potentially at higher risk of incorporation bias than longitudinal studies and we accounted for this when assessing studies for risk of bias. The risk of incorporation bias may be higher in cross-sectional studies than cohort studies because the same examiner may do both index tests and reference standard, or the participant may make the examiner aware of the results of the other test. Conversely cross-sectional studies are at lower risk of bias due to participant flow causing partial verification of the reference standard. We judged it too restrictive to exclude cross-sectional studies.

We excluded case-control studies because they are at high risk of bias. Furthermore, it is unlikely that GPs would be able to make a blinded clinical judgement for a participant in case-control studies since the study would recruit participants based on disease state (dementia, cognitive impairment, or normal).

Participants

We included studies if they recruited participants from primary care. We defined primary care as first-contact health care provided by a non-specialist general practitioner in a continuing-care office setting. We excluded studies if a consultation with a non-specialist occurred in hospital (including outpatients or emergency departments) because we judged this as unlikely to represent primary care in the sense that was relevant to the review objective.

We included studies if GPs made a clinical judgement about the presence of cognitive impairment or dementia. We excluded studies if GPs made a judgement about the presence of cognitive impairment or dementia in all people attending primary care, regardless of age, as we judged this to be akin to screening. We included articles exclusively with GPs as the primary care provider, rather than other allied professions (e.g. advanced nurse practitioners, physician assistant, etc.) to avoid heterogeneity in studies by professional role, because the training requirements for these different roles vary.

Index tests

We defined clinical judgement as being unaided by any additional test, investigation, or inquiry beyond that which is immediately available to the clinician (Blaeuer 2013; Body 2014; Di 2013). This would include history taking, history taking from an informant, and physical examination if appropriate. In everyday practice a GP forms a clinical judgement after an encounter with a patient, in which the GP typically accesses and reviews the medical record as part of their consultation. In contrast, diagnostic accuracy research in general practice may define clinical judgement in three ways.

Firstly, there may be a documented diagnosis of cognitive impairment or dementia in the medical records, but this is likely to reflect the process of documentation rather than clinical judgement (documented approach). In one study the number of patients registered as having dementia on GP surgery records increased by 9% when diagnostic coding was systematically audited (Russell 2013). Documented diagnoses of dementia in the medical record may not reflect the clinical judgement of a GP about the presence or absence of the target condition, and indeed the increase in the prevalence with consistent coding suggests that documentation may be systematically under-coded in the electronic medical record in routine practice.

Secondly, studies may define clinical judgement as an opinion based on knowledge of the patient and review of the medical notes, but not relating to a specific encounter with the patient (retrospective approach). However, studies using this definition are likely to be affected by differences in consulting behaviour of people with cognitive impairment or dementia compared to those without these problems (Chen 2014; Ramakers 2007; Russell 2013; Ydstebø 2015). Because cognitive status is associated with consultation behaviour it may also be associated with the implicit (i.e. not formally recorded) knowledge that a GP has about an individual when forming a retrospective clinical judgement.

Thirdly, studies may define clinical judgement as the impression formed by a clinician after consulting with a patient who has presented to a specific encounter with the doctor (prospective approach). The patient may be consulting with symptoms suggestive of possible dementia, but not necessarily, because it may be the consulting GP, or indeed a third party, who raises the possibility of cognitive problems. We considered this as the definition of clinical judgement that is most relevant to practice for this review, being contemporaneous with a specific encounter, and least likely to be subject to systematic bias in coding in the medical record.

We included studies that used the third prospective definition. To avoid an empty review, we also included studies that used the second retrospective definition so long as the GP's determination about cognitive status had taken place before any definitive diagnosis. We excluded studies that used the documented approach as we considered this did not reflect clinical judgement in a sense that was applicable to practice.

As described above, clinical judgement informs the Clinical pathway, with an associated risk of partial-verification bias. We included studies where some (but not necessarily all) participants underwent both the index test and reference standard, so long as at least some index test positives and index test negatives underwent the reference standard, and we accounted for partial verification as described in Assessment of methodological quality.

We included studies that allowed GPs to use additional cognitive tests to help determine the management of the patient after formulating and expressing their unaided clinical judgement, but we accounted for these additional tests as a source of heterogeneity as described in Assessment of methodological quality. We only evaluated accuracy of unaided clinical judgement in this review.

Target conditions

There were two target conditions for the review: all-cause dementia, and cognitive impairment due to dementia or mild cognitive impairment (MCI) (Gauthier 2006).

Experience in clinical general practice is that patients focus on the possibility of dementia rather than MCI when they present with concerns about impaired cognition, but inevitably physicians will diagnose MCI in some people who undergo evaluation for possible dementia. The second target condition includes both dementia and MCI, because it would be unusual for a GP to diagnose MCI, especially based on clinical judgement alone. If clinical judgement was sensitive for any cognitive impairment, then if the GP assessed the person as being cognitively normal it would be likely that the person had neither dementia nor MCI.

We excluded studies if they investigated the accuracy of clinical judgement for risk prediction of future dementia. There was no restriction to a particular stage or clinical severity of dementia.

Reference standards

To allow for a pragmatic and sensitive approach to study inclusion, we included studies that used several different reference standards (outlined below). We accounted for the interval between index test and reference test in Assessment of methodological quality. Importantly, different definitions of dementia identify people with a different spectrum of disease. The most notable differences between the definitions are that: ICD-10 and DSM-IV-TR require memory impairment in conjunction with impairment in other domains, whereas DSM–5 only requires impairment in one domain and does not require amnesia; that only ICD-10 requires impairment in emotional regulation or social interaction; and that DSM-IV-TR and ICD-10 require impairments of daily life to be

attributable to each of the impaired cognitive domains. Finally ICD-10 states that for a "confident" diagnosis the symptoms should be present for at least 6 months. ICD-10 tends to identify people who have more advanced dementia (Erkinjuntti 1997).

Dementia

We included studies that used a definition from DSM-III-R, DSM-IV-TR (American Psychiatric Association 2000), or ICD-10 (ICD 1993). Any version of the Diagnostic and Statistical Manual of Mental Disorders (DSM), including DSM-V, or International Classification of Diseases (ICD) would have been eligible. Studies were also included if they used Automated Geriatric Examination for Computer Assisted Taxonomy (AGECAT) (Copeland 1986), Cambridge Mental Disorders of the Elderly Examination (CAMDEX) (Roth 1986), Clinical Dementia Rating Scale (Hughes 1982), or structured interview for the diagnosis of dementia of the Alzheimer type, multi-infarct dementia and dementias of other aetiology according to ICD-10 and DSM-III-R (SIDAM) (Zaudig 1991) as these are well-validated methods of applying the aforementioned diagnostic criteria.

We included studies if they used expert specialist clinical judgement as the reference standard; we defined a specialist as a clinician with expertise in diagnosing and managing dementia, practising in a hospital or secondary care environment, with the professional status of geriatrician, psychiatrist, or neurologist. We included studies if they used longitudinal confirmation of the diagnosis of all-cause dementia in primary care, because we anticipated that some studies would only offer a specialist assessment to some participants. We operationalised longitudinal confirmation of the diagnosis in primary care as case-record review occurring at least 3 months after the index test diagnosis of dementia where study case-note reviewers did not identify any other alternative diagnosis. Stage of disease did not form part of the target condition definition, though we recognised many people who GPs correctly diagnosed as having dementia by unaided clinical judgement (true positives) might have an advanced stage of disease.

Although the target condition was all-cause dementia, studies that used an aetiological subtype definition were also eligible, these were: for Alzheimer's disease the NINCDS-ADRDA criteria (McKhann 1984; McKhann 2011); for vascular dementia the NINDS-AIREN criteria (Román 1993); for Lewy body dementia the Dementia with Lewy Body Consortium criteria (McKeith 1996; McKeith 2005); and for frontotemporal dementia the consensus criteria (Neary 1994).

Cognitive impairment

Cognitive impairment was a composite target condition encompassing dementia (as defined above) and MCI. Any recognised definition of MCI was eligible e.g. the original or revised criteria of Petersen (Petersen 1999; Petersen 2004), Winblad (Winblad 2004), or the NINCDS-ADRDA criteria (McKhann 2011). Whilst acknowledging that causes of cognitive impairment extend beyond dementia and MCI (e.g. head injury, delirium, neoplasm, and non-neurodegenerative processes such as functional cognitive disorders) these were not part of the target condition for the review. Therefore, if (for example) the index tests indicated cognitive impairment or dementia, and further evaluation demonstrated that the clinical problem was neoplasm instead, the test would be false positive.

Search methods for identification of studies

Electronic searches

We searched MEDLINE (Ovid SP); Embase (Ovid SP); Web of Science Core Collection, including the Science Citation Index and the Conference Proceedings Citation Index (Thomson Reuters Web of Science); PsycINFO (Ovid SP), and LILACS (BIREME) on 16 September 2021. See Appendix 1 for the sources searched and search strategies used. Where appropriate, we used controlled vocabulary such as MeSH terms (in MEDLINE) and EMTREE (in Embase) and other controlled vocabulary in other databases.

Search filters are collections of terms aimed at reducing the number needed to screen by filtering out irrelevant records and retaining only those that are relevant. We did not use search filters designed to retrieve diagnostic test accuracy studies as a method to restrict the search overall for the MEDLINE, Embase, PsycINFO or LILACs search, because available filters have not yet proved sensitive enough for systematic review searches (Whiting 2011b). We included a validated filter for primary care studies that optimises sensitivity and specificity (Gill 2014). We did not apply any language restriction to the electronic searches.

Searching other resources

In addition, we checked the reference lists of all relevant papers for additional studies. We also searched:

- Meta-analyses van Diagnostisch Onderzoek (MEDION database) (www.mediondatabase.nl);
- NIHR Dissemination Centre (which replaced DARE) (evidence.nihr.ac.uk/);
- Health Technology Assessment Database (HTA database) in the Cochrane Library (www.cochranelibrary.com); and
- Aggressive Research Intelligence Facility (ARIF database) (147.188.28.230/rmwp).

We also talked to experts and attempted to contact authors where necessary to obtain details of unpublished studies.

Data collection and analysis

Selection of studies

Two review authors independently screened the retrieved citations at the title and abstract stage, using Covidence to classify each citation as relevant, possibly relevant, or not relevant. We resolved conflicts in classification by discussion until we reached consensus. We obtained full-text articles for all citations classified as either possibly relevant or relevant. We classified articles that we excluded at the full-text stage using the following hierarchy.

1. Inappropriate participants: not primary care.
2. Inappropriate reference standard.
3. Inappropriate index test: not clinical judgement.
4. Inappropriate target condition.
5. Inappropriate study design (i.e. not a diagnostic test accuracy study e.g. a study reporting qualitative data, descriptive epidemiology, randomised trial, or survey).

We classified articles under the highest order reason for exclusion, so that if a study was not set in primary care, then we excluded it at level one, whereas if it was set in primary care but did not use

an appropriate index test then we excluded it at level three. We required only one reason to exclude a study.

We emailed authors of included studies when necessary to obtain paired data on sensitivity and specificity if the original article did not report these data clearly. Additional Table 1 shows the circumstances under which we contacted authors in the hope of obtaining relevant information on diagnostic accuracy.

We presented data from multiple papers referring to the same study under a 'primary reference' based on the study that provided most data to our review.

Data extraction and management

Two review authors separately extracted data directly into Review Manager 2020 based on the list required for Cochrane Reviews of diagnostic test accuracy: sampling, characteristics of participants and setting, index test, target condition, reference test, flow and timing, use of prior tests and comparator tests. We also extracted data relating to study level covariates of average age, proportion of women participants, average scores on any cognitive test, stage or severity of dementia, average educational attainment for participants, average age and experience of general practitioners performing index test, and proportion of male and female doctors. The two authors discussed the data they had extracted and reached consensus on any discordant data by referring to the original full-text articles.

Where necessary we attempted to contact authors of included primary studies to obtain missing or unclear information relating to covariates listed above and/or items on the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) checklist.

Two review authors separately extracted paired data on sensitivity and specificity from included studies separately into Review Manager. We resolved discrepancies by discussion informed by further review of the original manuscript until we reached consensus. We described studies that did not report paired sensitivity and specificity but did not include these in the meta-analysis.

Assessment of methodological quality

Two review authors used Review Manager to separately extract data to assess the risk of bias for each study, using the QUADAS-2 checklist (Whiting 2011a). We resolved any disagreements by discussion. For studies where only a sample were verified by the reference standard, we used the population undergoing both tests as the denominator for diagnostic accuracy, and the prevalence of cognitive impairment and dementia in the total sample were recorded separately where possible.

Statistical analysis and data synthesis

We used paired data on diagnostic accuracy to calculate the accuracy of the index test for diagnosing the two target conditions: cognitive impairment (including both MCI and all-cause dementia), and all-cause dementia. We calculated diagnostic accuracy with 95% confidence intervals separately for each target condition, in all studies with available data. We did the main meta-analysis in the studies at lowest risk of bias (one or fewer QUADAS-2 domains at high risk of bias).

We initially performed meta-analyses on pairs of sensitivity and specificity in Stata version 13 using metandi and xtmelogit. However, since metandi does not allow analysis with fewer than four studies, and xtmelogit does not produce output parameters for plotting a summary receiver operating characteristic (ROC) curve, we re-did the analysis to generate summary ROC curves using MetaDTA: Diagnostic Test Accuracy Meta-Analysis v 2.01 (Freeman 2019; Patel 2020). MetaDTA uses the bivariate random-effects model approach to jointly model sensitivity and specificity, and then derives the parameters for the hierarchical summary receiver operating characteristics (HSROC) model using equivalence equations, which are then used to plot a summary ROC curve (Chu 2006; Freeman 2019; Harbord 2007; Macaskill 2010; Patel 2020; Reitsma 2005). Plotting a summary ROC curve allows investigators to observe where study points are plotted, and how close they lie in relation to the summary curve, which allows a clear depiction of heterogeneity (Macaskill 2010).

We analysed different all-cause dementia diagnostic criteria together. We did not do meta-analyses by aetiological subtype of dementia, because it is unlikely that GPs would make an aetiological subtype diagnosis.

Investigations of heterogeneity

We designated three sources of heterogeneity in advance of the analyses as being important to investigate: the definition used to define the reference standard (ICD-10, DSM-III-R or DSM-IV-TR); whether the clinical judgement of the GP was based on a prospective opinion or a retrospective opinion; and whether the GP had access to use medical records when giving their opinion. We defined a further investigation of heterogeneity when the characteristics of the included studies were known, to investigate heterogeneity by risk of bias in the flow and timing QUADAS-2 domain (high risk or not). We selected this QUADAS-2 domain as being particularly important to investigate heterogeneity in given the setting and potential for partial verification bias. We initially investigated heterogeneity through visual examination of paired diagnostic accuracy data in forest plots, and the ROC plot of the raw data, and then by adding these components (definition, opinion, records, flow and timing) of study design as covariates to the analytical model. We used likelihood ratio tests to compare model fit (Macaskill 2010).

We specifically did not examine the length of vocational training programme for GP participants as a source of heterogeneity because we anticipated in advance this would be poorly reported (or not reported) in original studies. Furthermore, we followed recommendations to only consider possible sources of heterogeneity which vary at the study level (Bossuyt 2013), which is unlikely to be the case for these domains.

We advise caution when interpreting investigations of heterogeneity, especially exploratory post-hoc analyses (Bossuyt 2013).

Sensitivity analyses

We investigated how including studies that we judged to be at high risk of bias in more than two QUADAS-2 domains influenced the estimates of diagnostic accuracy. We had also planned to investigate the impact of excluding studies that used extended primary care follow-up or expert clinical judgement as the

reference standard, from the analysis, but in the event there were no studies in this category.

Assessment of reporting bias

Quantitative methods for exploring reporting bias are not well established for studies of diagnostic test accuracy (Bossuyt 2013) and so this was not examined.

RESULTS

Results of the search

Figure 1 shows the flow of studies in the review. We retrieved a total of 18,202 results and after de-duplication there were 12,427 records to assess.

Figure 1. Study flow diagram.

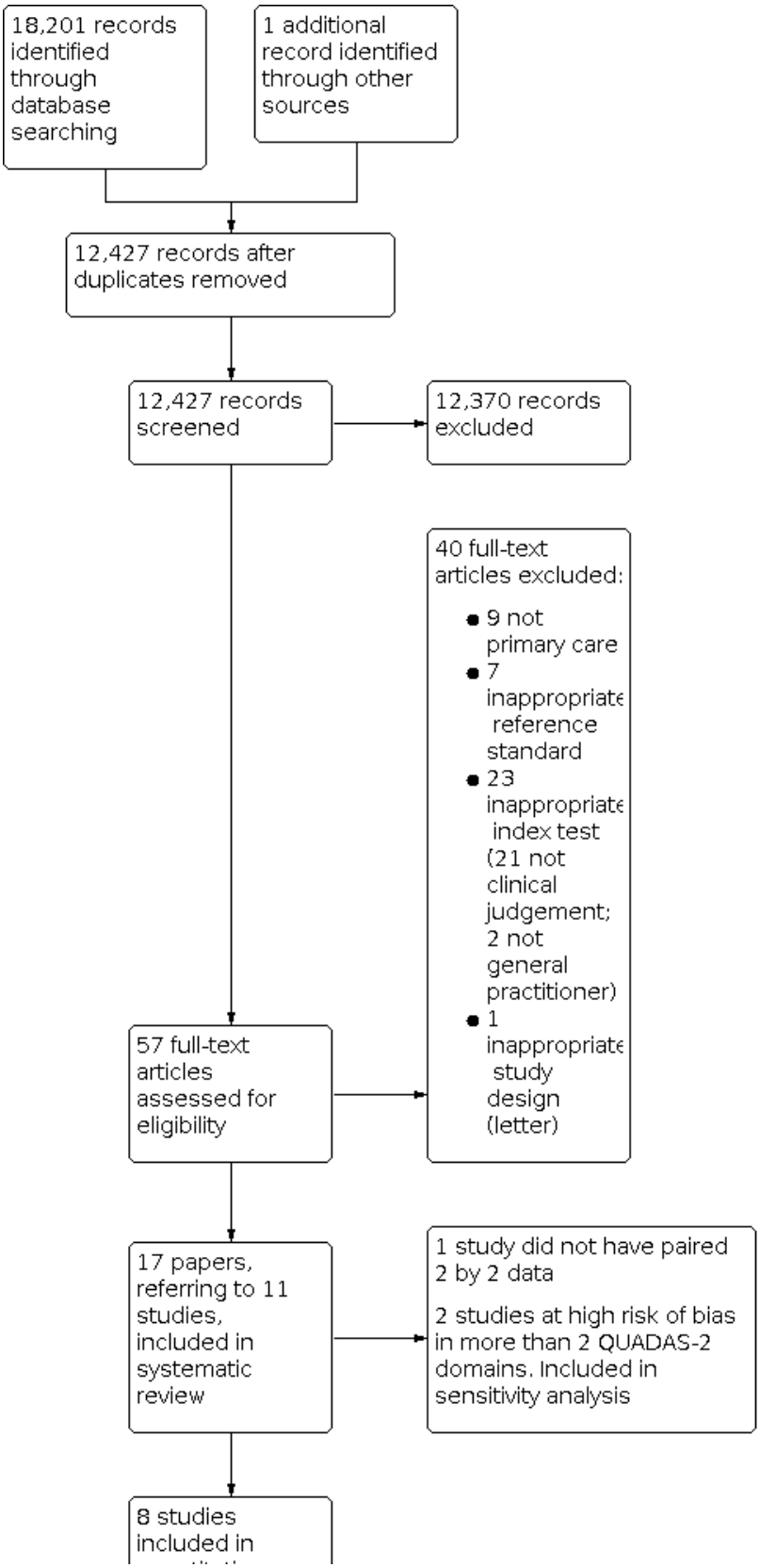
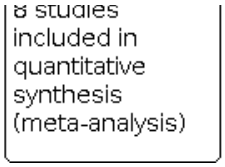


Figure 1. (Continued)



Review of the full text for 57 records led to us including 17 records, referring to 11 studies, of which we included eight in the meta-analysis. We included Valcour 2000 in the review but not in the meta-analysis because it was not possible to obtain paired data on diagnostic accuracy either from the original paper or after correspondence with the authors. We included the study for transparent reporting because based on the study design paired diagnostic accuracy data should be available. We did not include Pentzek 2009 or De Lepeleire 2004 in the main meta-analysis because Assessment of methodological quality indicated that they were at high risk of bias in two QUADAS-2 domains, but we did include these two studies in pre-planned Sensitivity analyses.

Of the 40 papers that we excluded at the full-text review stage, we excluded nine because they were not based in primary care. Instead, these papers were set in the general community, i.e. not people consulting in primary care - there was no consultation with a clinician (Borson 2006; Chong 2016; Engedal 1989;Fichter 1995; Pittmann 1992), outpatients department (Belmin 2012; Jacinto 2009; Wilkins 2007) or hospital (Dilts 2003). Seven papers were excluded because they used an ineligible reference standard; instead of using a reference standard from the pre-specified list, the excluded papers used methods which would be at high risk of incorrectly classifying the target condition. Of the seven papers excluded because of an ineligible reference standard, four used cognitive tests as the reference standard, of which three (Jansen 2007; Tierney 2014; Waldorff 2005) used the Mini-Mental State Examination (MMSE) (Folstein 1975) and one (Mant 1988) used the Blessed Dementia Scale (Blessed 1968), two (Bushnell 2004; Hara 2013) did not have dementia as a target condition and used either a screening test for mild cognitive impairment (MCI) which was not further detailed (Hara 2013) or the World Health Organization Composite International Diagnostic Interview (CIDI) (Bushnell 2004), and one aimed to validate a new measure and cross-validated the new tool against other cognitive tests but not a diagnosis (Hopman-Rock 2001).

We excluded one study, a letter, because it was not a diagnostic test accuracy study (Leung 2007).

Of the 23 papers that we excluded because they investigated an ineligible index test, two papers were not investigating the accuracy of general practitioners (GPs) (Livingston 1990; Schaub 2003). Of the remaining 21 papers which we excluded because of an ineligible index test, five (Aldus 2018; Camicioli 2000; Löppönen 2003; Mok 2004; Olafsdóttir 2000) investigated the documentation of a diagnosis in the medical record which was specified in advance as an ineligible index test. Of the remaining 16 papers that had an ineligible index test, eight papers (van Hout 1999; van Hout 2000; van Hout 2001; van Hout 2002; van Hout 2003; van Hout 2006; van Hout 2007a; van Hout 2007b) referring to one study were excluded because GPs were asked to use the Dutch dementia guidelines to make a diagnosis rather than their unaided clinical judgement; three papers (Dinesen 1997; Lionis 2001; Wang 2017) were excluded because GPs were asked to use the MMSE (Folstein 1975); one study (De Lepeleire 2005) used the ADMP scale (Heyrman 1990); one study (Hessler 2014) used the 6CIT (Brooke 1999); one study (Juva 1994) used the Clinical Dementia Rating (CDR) scale (Morris 1993); one study (Kurz 1999) used a list of warning signs of dementia from the Alzheimer's Association; and one study (Noda 2018) used the standardised physicians' manual issued by the Ministry of Health, Labor and Welfare of Japan.

We identified two previous systematic reviews (Mitchell 2011; van den Dungen 2012). Additional Table 2 compares the papers that were included in the two previous reviews with the current review and indicates that there were four studies that were included in all three reviews (Cooper 1992; Eefsting 1996; Pond 1994; Valcour 2000); three studies that were included only in the current review (Brayne 1990; Creavin 2021; De Lepeleire 2004); four studies that were included in both the current review and Mitchell 2011 (Kaduszkiewicz 2014; O'Connor 1993; Pentzek 2009; Rondeau 2008); two studies that were included in both Mitchell 2011 and van den Dungen 2012 but not in the current review (Löppönen 2003; Olafsdóttir 2000); and 12 studies that were only included by Mitchell 2011 (Boise 2004; Borson 2006; Boustani 2005; Bowers 1990; Callahan 1995; Chodosh 2004; Ganguli 2004; Iliffe 1990; Jacinto 2009; Mant 1988; van Hout 2000; Wilkins 2007).

Of the 14 studies that previous reviews included but we did not, we excluded 12 because the index test did not meet our inclusion criteria, and we excluded two because the reference standard did not meet our inclusion criteria (Bowers 1990; Mant 1988). Of the 12 studies that we excluded from the current review because the index test did not meet the inclusion criteria one study (van Hout 2000) required GPs to use the Dutch dementia guidelines to make a diagnosis of dementia and the other 11 studies defined clinical judgement as a documented diagnosis of dementia in the medical record, neither of which met the criteria for clinical judgement as defined in the current review.

Additional Table 3 summarises the characteristics of included studies regarding sampling, index test, reference standard, participant flow, target condition, definition, and access to medical records. Additional Table 4 provides a further overview of the participant selection, the characteristics of participants, verification of the target condition, and the proportion classified by GP judgement and the reference standard. The characteristics of Valcour 2000 are described but because there was no usable 2 x 2 data on diagnostic accuracy this study was not included in the meta-analysis. Eight studies sampled consecutive consulting patients or the entire registered GP list (Cooper 1992; Creavin 2021; De Lepeleire 2004; Eefsting 1996; O'Connor 1988; Pond 1994; Rondeau 2008; Valcour 2000) and three took a sample of patients from GP surgery lists (Brayne 1990; Pentzek 2009; Wind 1995). Four studies (Brayne 1990; Eefsting 1996; O'Connor 1988; Wind 1995) used a retrospective judgement, and seven (Cooper 1992; Creavin 2021; De Lepeleire 2004; Pentzek 2009; Pond 1994; Rondeau 2008;

Valcour 2000) used a prospective judgement. Six studies (Brayne 1990; Cooper 1992; De Lepeleire 2004; Eefsting 1996; O'Connor 1988; Wind 1995) used CAMDEX as the reference standard, one used the Canberra Interview for the Elderly (Pond 1994) or the SIDAM (Pentzek 2009) and three (Creavin 2021; Rondeau 2008; Valcour 2000) used expert opinion. Six studies (Brayne 1990; Creavin 2021; O'Connor 1988; Pentzek 2009; Valcour 2000; Wind 1995) appeared to have complete verification of the index test, with the reference standard being administered to all participants, but only three of these (Brayne 1990; Creavin 2021; Pentzek 2009) did not screen people in some way prior to the index test; the other studies had some form of partial verification. All studies investigated dementia as a target condition and five studies (Cooper 1992; Creavin 2021; Eefsting 1996; Pentzek 2009; Pond 1994) additionally investigated MCI as a target condition. Six studies (Brayne 1990; Cooper 1992; Creavin 2021; O'Connor 1988; Pond 1994; Wind 1995) used ICD-10 as the definition, two studies (De Lepeleire 2004; Pentzek 2009) used DSM-IV-TR, and one study used each of DSM-III-R (Eefsting 1996), CDR (Valcour 2000), and NINCDS-ADRDA (Rondeau 2008). Access to the medical records was not available in two studies (Brayne 1990; Cooper 1992), was unclear in one (O'Connor 1988), and was available in the remainder.

Methodological quality of included studies

We used QUADAS-2 to judge the risk of bias in each study. Figure 2 shows that overall there was low risk of bias in the included studies. Specifically most studies were at low risk of bias in the patient selection domain, with the remainder being at unclear risk of bias; and there was low concern about applicability. All studies were at low risk of bias in the index test domain and most studies had low concern about applicability. Most studies were at low risk of bias in the reference standard domain and all had low concern about applicability, though two were at high risk of bias. Flow and timing was the domain where there was most risk of bias. Figure 3 presents the summary risk of bias and applicability assessment for individual studies. Four studies were at unclear risk of bias in the patient selection domain. Pentzek 2009 required potential participants to have had at least one GP contact in the past 12 months and excluded people who were housebound, which may have resulted in the systematic exclusion of people who lived alone or did not seek help, or with the most severe impairment. Similarly De Lepeleire 2004 specifically excluded people who lived in a residential home for the elderly (regardless of whether home visits were needed) and O'Connor 1988 excluded people who lived in long stay hospitals, and also those who were diagnosed as having "minimal dementia", which was not defined. Rondeau 2008 stated that consecutive patients were enrolled, but it was not clear that these patients were genuinely consecutive because five patients per GP were recruited over 2 years, which appears low. There were no concerns about the applicability of patient selection.

Figure 2. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies.

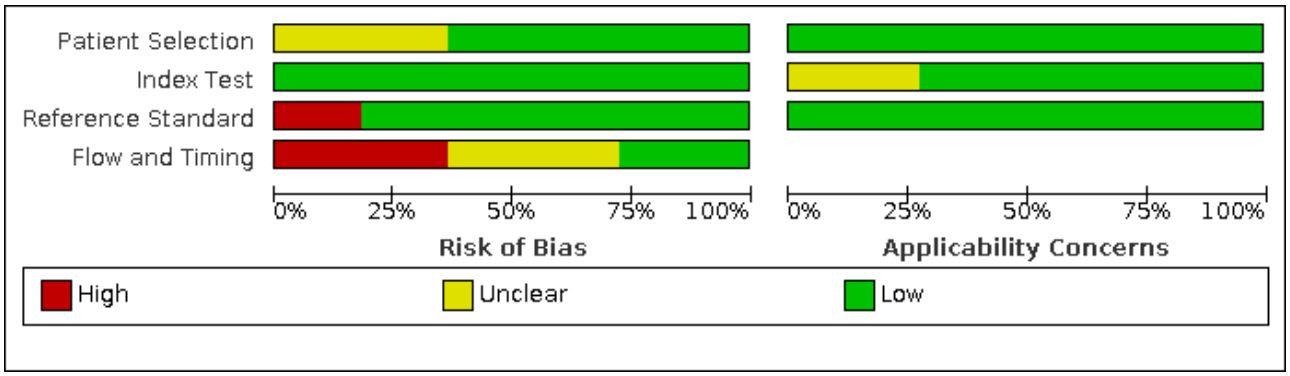
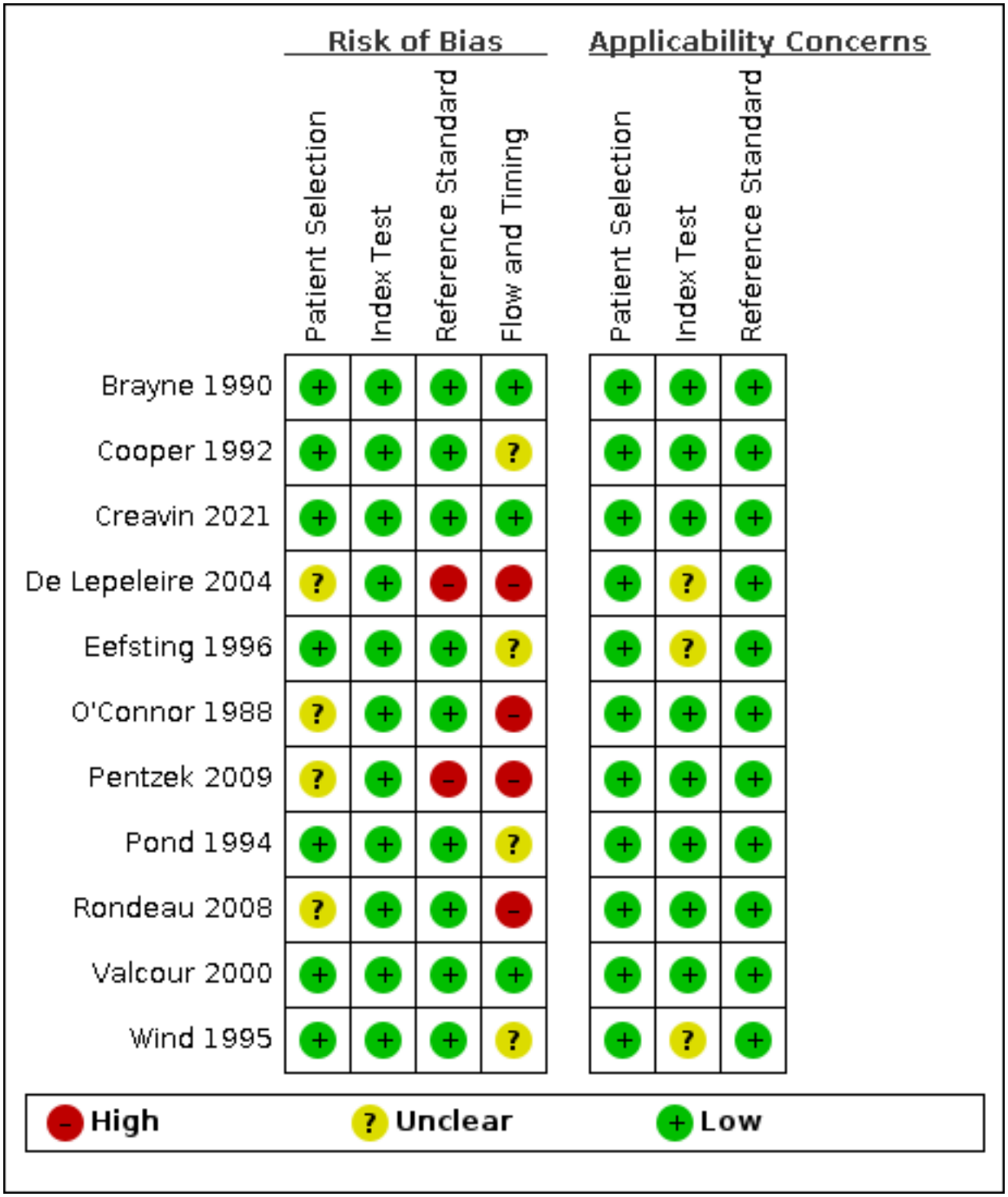


Figure 3. Risk of bias and applicability concerns summary: review authors' judgements about each domain for each included study.



We judged all studies to be at low risk of bias in the index test domain. However, three studies had unclear applicability of the index test. Wind 1995 and Eefsting 1996 explained the reference standard criteria (but not the diagnosis) to participating GPs before they made their judgement. De Lepeleire 2004 asked GPs to give their opinion after they had asked four questions on a four-item test of instrumental activities of daily living (Barberger-Gateau 1999), though did not specifically state the GPs had to use the results of this instrument.

We judged two studies to be at high risk of bias in the reference standard domain. We judged both Pentzek 2009 and De Lepeleire 2004 to be at high risk of bias in this domain because they incorporated the index test into the reference standard. No studies

were at unclear risk of bias in the reference standard domain and there were no concerns about the applicability of the reference standard.

We judged four studies to be at high risk of bias in the flow and timing domain. Studies at high risk of bias in the flow and timing domain are generally at risk of partial verification, and may not have fully ascertained all cases of people with dementia. Pentzek 2009 applied the reference standard at 1.5 years and 3 years after the index test assessment. O'Connor 1988 did not provide any information on timing, but was at risk of partial verification because the reference test CAMDEX was used to evaluate people who scored 23 or less on the MMSE, together with sample of those who scored 24 or 25, but none of those who scored over 25. Rondeau 2008 was also at high risk of bias in this domain because of partial verification: 222 of 375 (59%) people diagnosed with dementia by a GP were seen by a specialist, in contrast 38 of 711 (5%) people not diagnosed with dementia by a GP were seen by a specialist, and 125 of 311 (40%) people with an uncertain GP diagnosis were seen by a specialist. De Lepeleire 2004 was at high risk of bias in this domain because it appeared that the reference standard was only done on 10 people of 1003 who were evaluated by a GP. Cooper 1992 was at unclear risk of bias because there was partial verification but a stratified random sample was taken for verification, containing equal numbers in each of the four categories of GP assigned impairment. Similarly, Pond 1994 took a random sample for verification by the reference standard and was at unclear risk of bias. Wind 1995 was at unclear risk of bias because although there appeared to be full verification, this was derived from people who had been screened with the MMSE prior to the index test, including a sample of those who scored up to 30, and there was no information on timing. Eefsting 1996 performed the reference standard on all participants scoring 17 or below on the MMSE, together with a random 2/3 sample of those scoring 18 to 23, a random 1/3 sample of those scoring between 24 and 27 and none of those scoring 28 and above, and was at unclear risk of bias.

Two studies, Pentzek 2009 and De Lepeleire 2004, were judged to be at high risk of bias in two QUADAS-2 domains (reference standard, and flow and timing).

Findings

Two studies reported the prevalence of dementia, Eefsting 1996 reported a prevalence of 7% and O'Connor 1988 reported a prevalence of 11%. In both cases the reported prevalence takes account of weighting in the sampling for the reference test and so differs from the raw calculation of true positive (TP) + false negative (FN)/(TP + FN + false positive (FP) + true negative (TN)).

Figure 4 is a forest plot of the accuracy of clinical judgement of GPs for the diagnosis of dementia in the 10 studies that had paired data on sensitivity and specificity. Excluding De Lepeleire 2004 , which reported a sensitivity and specificity of 100% but was also one of two studies that were at high risk of bias in two QUADAS-2 domains, in individual studies sensitivity ranged from 34% in Brayne 1990 to 91% in Cooper 1992, which was one of only two studies (the other being Rondeau 2008) that reported higher sensitivity than specificity. Specificity was generally higher than sensitivity and ranged from 58% in Rondeau 2008 to 99% in Eefsting 1996.

Figure 4. Forest plot of clinical judgement for dementia.

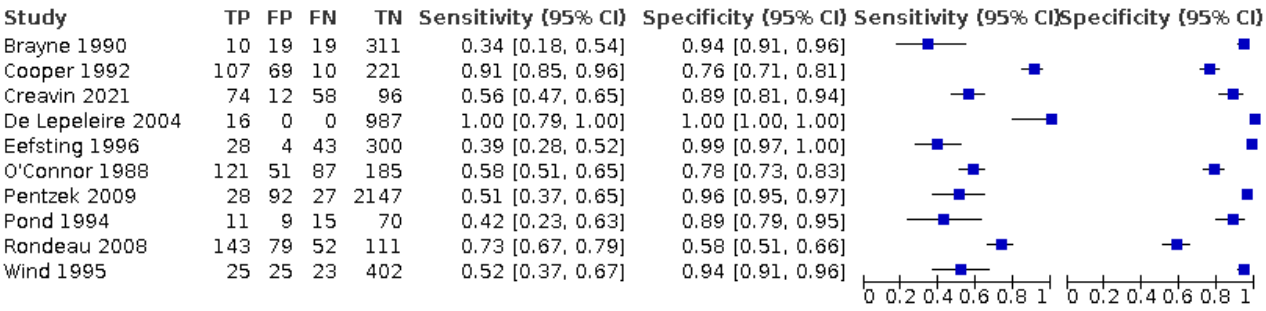


Figure 5 is a summary plot of the accuracy of clinical judgement of GPs for the diagnosis of dementia. The dashed bubble indicates the 95% confidence interval (CI) around the summary point and the larger dotted bubble indicates the 95% prediction region. The 95% CI indicates the region where in 95 samples out of 100 the calculated 95% CI will contain the true mean value, based on the included data (Takwoingi 2015). The 95% prediction region indicates the area where the results of a future study could be expected to lie, based on the analysed data (Takwoingi 2015). In the meta-analysis for dementia as the target condition, the summary diagnostic accuracy of clinical judgement of general practitioners was sensitivity 58% (95% CI 43% to 72%), specificity 89% (95% CI 79% to 95%), diagnostic odds ratio (DOR) 11 (95% CI 4.8 to 18), positive likelihood ratio 5.3 (95% CI 2.4 to 8.2) and negative likelihood ratio 0.47 (95% CI 0.33 to 0.61). As shown in Figure 5 the summary point is an average of the studies in the meta-analysis. Only four of the eight studies in the meta-analysis were included in the 95% confidence region displayed on the summary receiver operating characteristic (ROC) curve, with one lying on the edge of the 95% confidence region (Creavin 2021), one just outside (Eefsting 1996), and two lying further outside (Cooper 1992; Rondeau 2008). Therefore the summary point in isolation is over-simplistic as a representation of the diagnostic accuracy of clinical judgement for dementia and does not reflect the heterogeneity in the data.

Figure 5. Summary plot of general practitioners' clinical judgement for diagnosis of dementia.
Colours of dots: indicate risk of bias in flow and timing domain. Surround indicates prospective (blue) retrospective (yellow); decimals indicate prevalence of target condition.

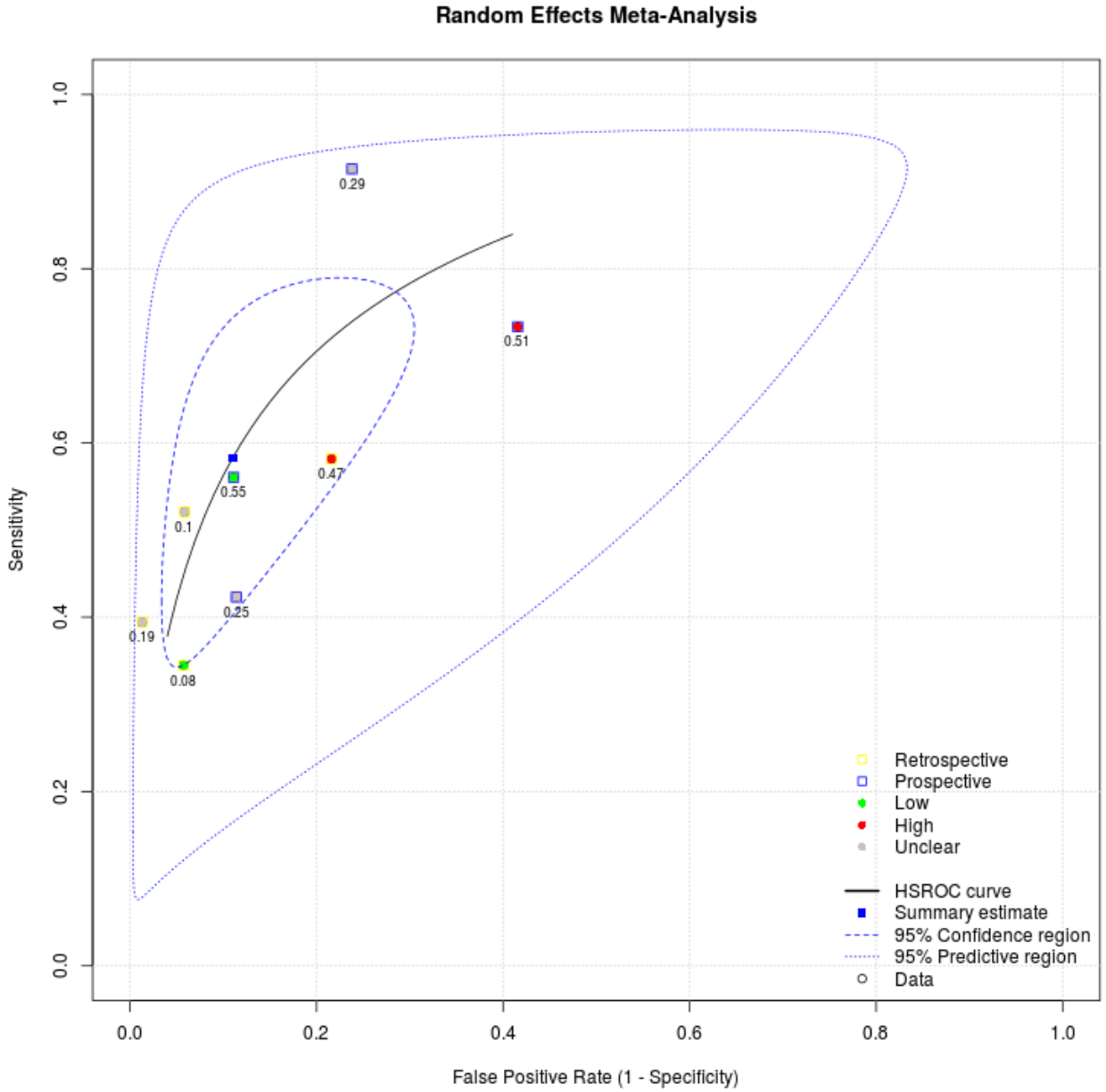


Figure 6 is a forest plot of the accuracy of clinical judgement of GPs for the diagnosis of cognitive impairment in the four studies that had paired data on sensitivity and specificity. Excluding Pentzek 2009, which was at high risk of bias in two QUADAS-2 domains and reported sensitivity 12% (95% CI 9% to 15%) specificity 94% (95% CI 93% to 95%), in individual studies sensitivity ranged from 58% in Wind 1995 to 97% in Cooper 1992, which was the only study that reported higher sensitivity than specificity. Specificity ranged from 40% in Creavin 2021 to 88% in both Wind 1995 and Eefsting 1996.

Figure 6. Forest plot of clinical judgement for cognitive impairment.

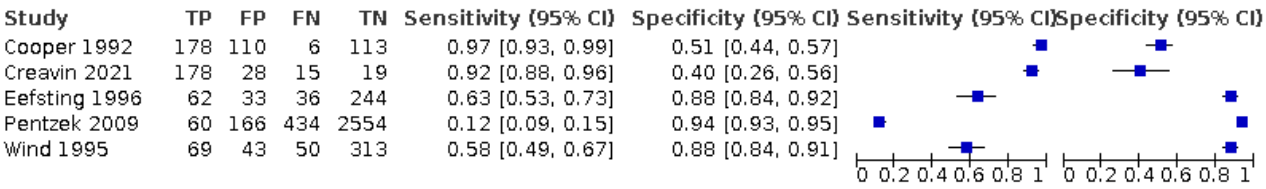
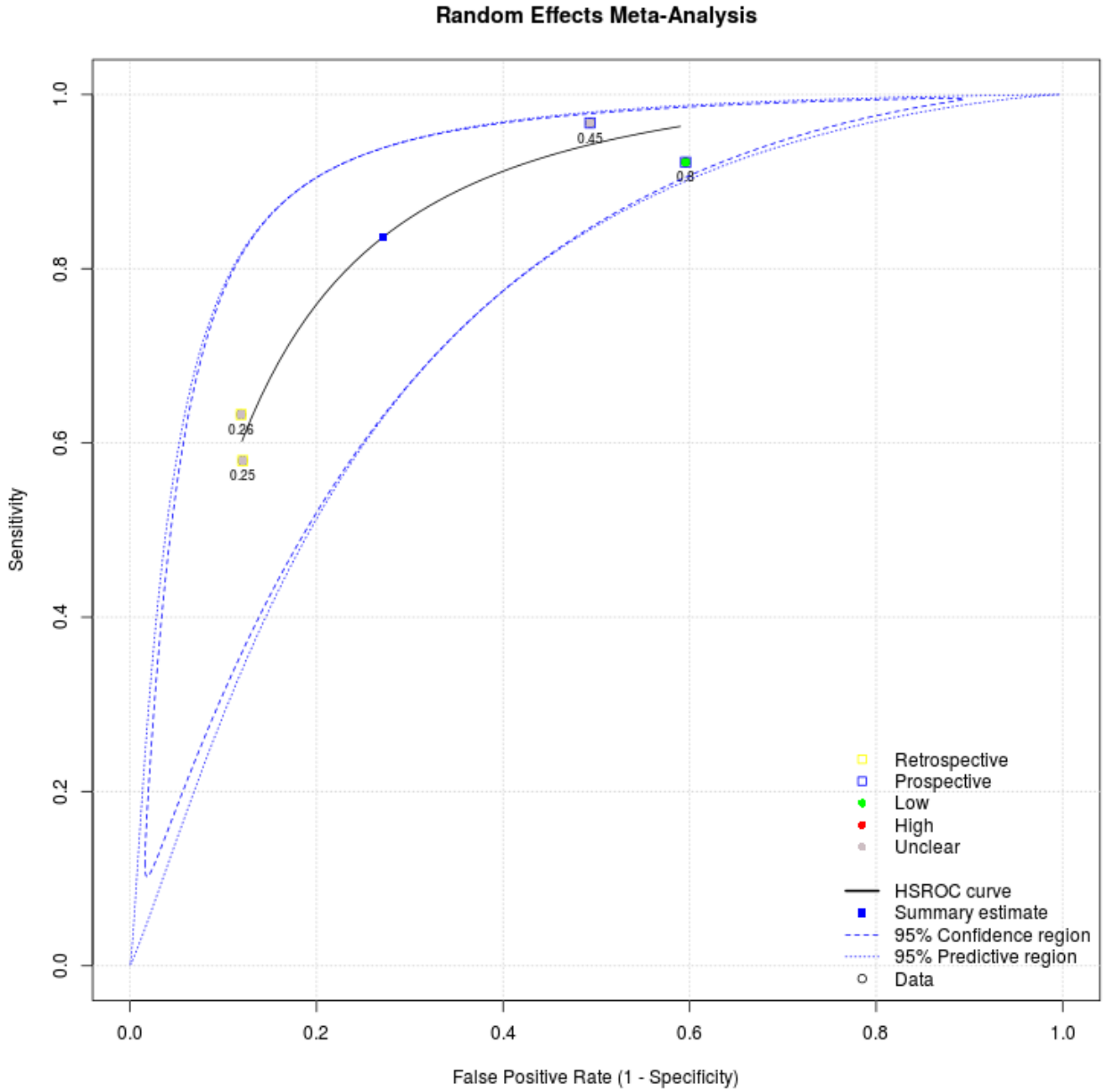


Figure 7 is a summary plot of the accuracy of clinical judgement of GPs for the diagnosis of cognitive impairment. In the meta-analysis for cognitive impairment (including dementia) as the target condition, the summary diagnostic accuracy of clinical judgement of general practitioners was sensitivity 84% (95% CI 60% to 95%), specificity 73% (95% CI 50% to 88%), DOR 14 (95% CI 8.4 to 19), positive likelihood ratio 3.1 (95% CI 1.4 to 4.7) and negative likelihood ratio 0.23 (95% CI 0.06 to 0.40). No single study had a comparable diagnostic accuracy to the summary point on either sensitivity or specificity. Furthermore, the 95% confidence region was the same size as the 95% prediction region which covered a large amount of ROC space, and indicates a high level of uncertainty in the results. Therefore the summary point in isolation is over-simplistic as a representation of the diagnostic accuracy of clinical judgement for cognitive impairment and does not reflect the heterogeneity in the data.

Figure 7. Summary plot of general practitioners' clinical judgement for diagnosis of cognitive impairment. Colours of dots: indicate risk of bias in flow and timing domain, all but 1 are at unclear risk of bias in this domain. Surround indicates prospective (blue) retrospective (yellow); decimals indicate prevalence of target condition.



Heterogeneity

Unfortunately due to technical limitations described in Methods it was not possible to plot all of the data with two curves on a single ROC space. Many of the summary ROC plots had significant uncertainty and covered a large amount of ROC space, and so we describe our findings below rather than present ROC plots. All plots had significant uncertainty in the estimates. Because of this, it is impossible to quantify the extent to which the four sources of heterogeneity account for all of the heterogeneity in the test accuracy, and the extent to which heterogeneity in test accuracy is driven by other factors.

Definition

We plotted a summary ROC plot for the studies by definition, with one plot for studies that used the ICD-10 definition and one plot for studies that used either DSM-III-R or DSM-IV-TR. It is possible there was some heterogeneity in specificity by definition, with studies that used the ICD-10 or DSM-III-R definition appearing to report a higher specificity, but this could be due to chance. The study that used the DSM-IV-TR definition (Rondeau 2008) had a higher sensitivity than the two studies that used a DSM-III-R definition (Eefsting 1996; Pond 1994) at a cost of lower specificity, but this finding could be due to chance. Alternatively, the specificity in

Rondeau 2008 could be an outlier, possibly related to the high risk of bias in the flow and timing domain.

Index test

We plotted a summary ROC plot for the studies by index test, with one summary plot for the studies that used a prospective clinical judgement and one summary ROC plot for the studies that used a retrospective clinical judgement. There was no strong visual evidence to support the possibility of heterogeneity in specificity between prospective and retrospective judgement, but it was possible that a retrospective definition may be more specific than a prospective definition. Similarly, there was no strong visual evidence to support the possibility of heterogeneity in sensitivity between prospective and retrospective judgement, but was possible that a prospective judgement has a higher sensitivity than a retrospective judgement, but this could be due to chance.

Medical records

We plotted a summary ROC plot for the studies by access to the medical records, with one summary plot for the studies that allowed access to the medical records and one summary ROC plot for the studies that did not definitely allow access. It is plausible that studies that reported access to the medical records was available had higher sensitivity than those where access was not available or uncertain, however, there are two studies where access was available where sensitivity was low (Eefsting 1996; Pond 1994) so this finding could be due to chance, especially as one of the two studies with higher sensitivity (Rondeau 2008) was at high risk of bias in the flow and timing QUADAS-2 domain. There was no strong visual evidence to support the possibility of heterogeneity in specificity by access to the medical records.

Flow and timing

We plotted a summary ROC plot for the studies by risk of bias in the QUADAS-2 flow and timing domain, with one summary plot for studies at low and unclear risk of bias and one summary ROC plot for studies at high risk of bias in the flow and timing domain. Removing the studies at high risk of bias in the flow and timing QUADAS-2 domain reduced the uncertainty. However, the 95% prediction interval still contained a large amount of ROC space. It is plausible that studies that were at high risk of bias in the flow and timing domain had lower specificity than studies that were at low or unclear risk of bias in this domain. However, this finding could also be due to chance, because only two studies were at low risk of bias in the flow and timing domain (Brayne 1990; Creavin 2021).

Cognitive impairment target condition

There were four studies in the meta-analysis with cognitive impairment as the target condition, which restricted the opportunity to explore heterogeneity. For definition, three studies used the ICD-10 definition and one used the DSM-III-R definition. For index test, two studies used a retrospective judgement and two used a prospective judgement. For access to the medical records, all studies allowed access. For risk of bias in the flow and timing QUADAS-2 domain, three studies were at unclear risk of bias and one was at low risk of bias. Two studies used a prospective index test, and both used the ICD-10 definition for the target condition (Cooper 1992; Creavin 2021); whereas two studies used a retrospective index test, of which one used the ICD-10 definition for the target condition (Wind 1995) and one used the DSM-III-R definition (Eefsting 1996). There was no evidence of heterogeneity by definition for the cognitive impairment target condition, because two studies that used different definitions (Eefsting 1996, DSM-III-R; Wind 1995, ICD-10) had similar sensitivity and specificity. However, one of the two studies (Cooper 1992) that used the ICD-10 definition had higher sensitivity and lower specificity than Eefsting 1996 so it is impossible to draw firm conclusions. Because of the small number of studies there is a possibility of a type 2 error, i.e. that there may be heterogeneity which was not evidenced. There is a possibility that prospective clinical judgement is more sensitive and less specific than retrospective clinical judgement, but there is significant uncertainty and no firm conclusions can be drawn. There were insufficient studies to perform meta-regression.

Sensitivity analyses

Dementia target condition

We plotted a summary ROC plot for the pre-specified sensitivity analysis which included the two studies (De Lepeleire 2004; Pentzek 2009) that we judged to be at high risk of bias in two QUADAS-2 domains. Compared to the main analysis there was greater uncertainty and more heterogeneity in the data. One of the studies that we excluded from the main analysis (De Lepeleire 2004) was an outlier, with very high sensitivity and high specificity; the other study at high risk of bias in two QUADAS-2 domains (Pentzek 2009) had more comparable sensitivity and specificity to the other eight studies. The summary estimate was sensitivity 63% (95% CI 46% to 78%) specificity 94% (95% CI 83% to 98%). This compares to sensitivity 58% (95% CI 43% to 72%), specificity 89% (95% CI 79% to 95%) in the main analysis. Therefore, when including the two studies that were judged to be at high risk of bias in more than two QUADAS-2 domains, the sensitivity was 5 percentage points higher and the specificity was 5 percentage points higher; both were within the 95% CI for the main analysis. However, the uncertainty in the estimates was increased substantially as shown by the wider 95% confidence and prediction regions in the summary ROC plot and because of heterogeneity the point estimates are not a good representation of the data.

Cognitive impairment target condition

We plotted a summary ROC plot for the pre-specified sensitivity analysis for the target condition cognitive impairment. Compared to the main analysis there was greater uncertainty and more heterogeneity in the data. The study that we excluded from the main analysis (Pentzek 2009) was an outlier, with very low sensitivity compared to the other three studies. The summary estimate was sensitivity 72% (95% CI 33% to 93%) specificity 79% (95% CI 57% to 91%). This compares to sensitivity 84% (95% CI 60% to 95%), specificity 73% (95% CI 50% to 88%) in the main analysis. Therefore, when including the study that was judged to be at high risk of bias in more than two QUADAS-2 domains, the sensitivity was 12 percentage points lower and the specificity was 6 percentage points higher; both were within the 95% CI for the main analysis. However, the uncertainty in the estimates was increased substantially as shown by the wider 95% confidence and prediction regions in the summary ROC plot and because of heterogeneity the point estimates are not a good representation of the data.

DISCUSSION

Summary of main results

We searched five electronic databases for studies investigating the accuracy of clinical judgement of general practitioners (GPs) for the diagnosis of two target conditions, dementia, and cognitive impairment including dementia. From 12,427 records retrieved, we included 17 articles referring to 11 studies in the review. Of these 11 studies, one study did not report paired data on test accuracy and two were at high risk of bias in two QUADAS-2 domains, leaving eight for inclusion in the meta-analysis for dementia as the target condition and four for cognitive impairment as the target condition. Overall we judged the included studies to be at low risk of bias and to have low concern about applicability. However, four of the eight studies included in the meta-analysis were at high risk of bias in the flow and timing QUADAS-2 domain.

We summarised our findings in Summary of findings 1. In the eight studies in the meta-analysis for dementia as the target condition, the test accuracy ranged from sensitivity 34% (95% confidence interval (CI) 18% to 54%) for Brayne 1990 to 91% (95% CI 85% to 96%) for Cooper 1992; and specificity ranged from 58% (95% CI 51% to 66%) for Rondeau 2008 to 99% (95% CI 97% to 100%) for Eefsting 1996. In the four studies in the meta-analysis for cognitive impairment as the target condition, the test accuracy ranged from sensitivity 58% (95% CI 49% to 67%) for Wind 1995 to 97% (95% CI 93% to 99%) for Cooper 1992; and specificity ranged from 40% (95% CI 26% to 56%) in Creavin 2021 to 88% (95% CI 84% to 91%) for Wind 1995 and 88% (95% CI 84% to 92%) for Eefsting 1996. Because of the wide 95% confidence and prediction intervals there is significant uncertainty in the findings and they should not be regarded as definitive. Cooper 1992 had very high sensitivity for dementia, which appears an outlier. This may be attributable to the sampling process in this study which took a stratified sample for verification potentially resulting in some people with relatively mild dementia not being identified by the reference standard and incorrectly labelled as true (rather than false) negatives, hence increasing sensitivity, or this may be or indeed be a true chance finding. Cooper 1992 also had relatively low specificity and this may be related either to the four-level classification process that GPs used to label people as having dementia, which may have helped GPs to conceptualise dementia as a spectrum rather than only identifying people with florid disease, alternatively the relatively low specificity in Cooper 1992 may be an artefact of the stratified sampling process, or indeed a true chance finding.

There was heterogeneity in the data between studies, as illustrated in Figure 5 and Figure 7. Diagnostic accuracy reviews often find greater heterogeneity in sensitivity than specificity because primary diagnostic accuracy studies typically contain fewer people with disease than without disease (Bossuyt 2013). It was difficult to draw firm conclusions about heterogeneity but the data were compatible with studies that used ICD-10, or applied retrospective judgement, having higher specificity compared to studies with DSM definitions or using prospective judgement. To the extent that there is genuinely heterogeneity by reference standard, this may be attributable to differences in reference standard described in Reference standards. For sensitivity there was no evidence of heterogeneity by definition, studies that used a prospective index test may have had higher sensitivity than studies that used a retrospective index test, and studies that allowed access to the medical records may have had higher sensitivity than studies that did not or where this was unclear. Studies at high risk of bias in the flow and timing domain appeared to have lower specificity and sensitivity than studies at unclear or low risk of bias in this domain. However, all of these findings could have been due to chance and there was significant uncertainty. We did consider exploring (post-hoc) heterogeneity by prevalence of the target condition as a proxy for disease severity. However, we judged that on balance an analysis of heterogeneity by derived prevalence may be misleading because we had concerns about the risk of bias in the flow and timing QUADAS-2 domain for many of the studies.

Strengths and weaknesses of the review

One limitation of the included studies, which is inherent in the general practice setting, is the relatively low prevalence of the target condition, which consequently leads to higher uncertainty in the estimates of diagnostic test accuracy and contributes to the heterogeneity between studies. With a prevalence of 30% two people will not have disease for every one person that does. The eight studies in the meta-analysis for the dementia target condition incorporated a total of 519 people who were true positives and 307 false negatives for the target condition dementia, totaling 826 people who were disease positive out of 2790, for a prevalence of 30% (95% CI 28% to 31%). In contrast the average prevalence of dementia in the included studies, calculated for each study as (true positives + false negatives)/total was 21%, and ranged from 2% in total (Pentzek 2009 and De Lepeleire 2004) to 47% in O'Connor 1988, 51% in Rondeau 2008 and 55% in Creavin 2021. Participant flow in studies, and specifically incomplete verification of the target condition with the reference standard means that calculated prevalence may not accurately reflect the true prevalence of the target condition. However, only two studies reported a figure for prevalence that accounted for the design of participant flow: Eefsting 1996 reported a prevalence of 7% and O'Connor 1988 reported a prevalence of 11%. In contrast, prevalence of dementia in the community is reported to be around 6% (Matthews 2013), but the prevalence of dementia in people attending health care would generally be expected to be higher than the general community since people attending health care are consulting about some problem.

The strength of evidence in this review is restricted by limitations in the primary studies regarding participant flow, and heterogeneity in the data. Only two studies in the meta-analysis (Brayne 1990; Creavin 2021) were judged to be at low risk of bias in the flow and timing QUADAS-2 domain. This may in part reflect the historic practical difficulties of investigating a disease in a low prevalence setting: large numbers of patients require evaluation to identify the people with disease; evaluating large numbers of people with a reference standard is resource intensive (i.e. expensive) and arguably burdensome for people who are unlikely to have a cognitive disorder. Partial verification may have led to over optimistic estimates of diagnostic accuracy in some studies. In contrast we judged all studies at low risk of bias in the index test QUADAS-2 domain.

In contrast, the evidence in this review is supported by methodological strengths of the included studies, such as participant sampling and the conduct of the index test and reference standard. We judged six of the eight studies in the meta-analysis to be at low risk of bias in the patient selection QUADAS-2 domain and the remainder to be at unclear risk of bias; all studies reported that participants were sampled either consecutively or

randomly. All eight studies were at low risk of bias in the reference standard QUADAS-2 domain.

The estimates of accuracy varied between studies but despite this there were consistent trends. In general, specificity for dementia was higher than sensitivity, specificity for dementia was relatively high (at least 0.78 in eight studies), whereas sensitivity for dementia was modest (at best 0.91 in eight studies). The findings were robust to a sensitivity analysis. Both prospective and retrospective clinical judgement for dementia and cognitive impairment generally had higher specificity than sensitivity. GPs missed fewer people when using the target condition cognitive impairment than when using the target condition dementia. These findings suggest that clinical judgement of GPs misses many cases of dementia (due to the low sensitivity) but if a GP judgement is of dementia it is likely this is correct, since the high specificity corresponds to a relatively low false positive rate: there is a good chance that someone is cognitively impaired if a GP thinks they have dementia. GPs are slightly better at identifying cognitive impairment (sensitivity 84% (95% CI 60% to 95%)) compared to dementia (sensitivity 58% (95% CI 43% to 72%)) for dementia, but the specificity of a GP diagnosis for cognitive impairment (73% (95% CI 50% to 88%)) is slightly lower compared to that for dementia (89% (95% CI 79% to 95%)).

There was low concern about the applicability of studies to the review question. Only three studies had unclear concern about the applicability of the index test but otherwise all studies were of low concern in all QUADAS-2 applicability domains.

Overall the characteristics of studies, quantity, numbers of disease positive, and consistency of the findings are supportive factors for the strength of evidence in the review. In contrast the strength of evidence is diluted by the heterogeneity in the data. However, heterogeneity is present in all diagnostic test accuracy reviews to some extent, and the nature of clinical judgement is that it will tend to be more heterogeneous than a machine read test.

Strengths and weaknesses of the review process

The search was comprehensive and systematic. A highly experienced information scientist, with substantial expertise in diagnostic test accuracy studies investigating cognitive disorders and dementia, wrote the search strategy. We searched several databases. Our search strategy identified all studies that were identified by previous systematic reviews, though we did not include all those studies in the review. The comprehensiveness of the search makes it unlikely that an unidentified study would undermine our findings, but we cannot exclude this possibility. In particular, it is hard to identify unpublished studies, and methods to identify reporting bias in diagnostic test accuracy reviews are not yet well established (Bossuyt 2013).

We excluded studies that used a documented diagnosis of dementia in the medical record as the index test. This meant that we excluded some studies that previous systematic reviews had included. Furthermore, it is not possible to infer the accuracy of documented diagnosis in the GP medical record for the diagnosis of dementia from this review. However, this exclusion means that the findings of this review are highly applicable to the review question which is arguably more focused on clinical practice than previous reviews. Understanding the accuracy of a documented diagnosis of dementia in the medical record would be helpful if (for instance) investigators were seeking to ascertain people with dementia diagnosed by a GP in a database of routinely collected data from primary care. Other investigators have reported inconsistency in ascertaining cases of dementia from routinely collected data (Sibbett 2017). We did not restrict our review to people with symptoms of dementia, and only we included two studies that had people with symptoms. It is arguable that studies that included people regardless of cognitive symptoms would be similar to screening studies, but there are two important caveats. Firstly, in contrast to standardised screening tests, although a clinical judgement is reached unavoidably in an encounter with a patient, it is critically dependent on the availability heuristic (Kahneman 2011) , that is, how much the diagnosis is in the mind of the consulting clinician, which would typically be increased if the clinician knew they were participating in a research study about dementia. Secondly, it is common for clinicians to opportunistically identify incidental pathology during consultations, even if the patient has not noticed a problem, though this is dependent on the availability heuristic and the severity of the problem.

The process for including and excluding studies was robust, as was data extraction, as we did this separately in duplicate by two review authors and resolved disagreements by consensus. We contacted authors of original studies when necessary to obtain data on diagnostic accuracy, but found these were not available. We used standard techniques that are recommended by the Cochrane diagnostic test accuracy review methods group for the statistical analysis. Recently reported approaches such as imputation of data (Ensor 2018) were not used, but are less widely recognised and are not yet mentioned in the Cochrane handbook (Bossuyt 2013).

Applicability of findings to the review question

Only three studies (De Lepeleire 2004; Eefsting 1996; Wind 1995) were of unclear applicability in the index test domain, whereas the other eight studies in the meta-analysis for dementia as the target condition were all of low concern about applicability. Overall there is low concern about applicability for the studies in the meta-analysis. The findings are applicable to addressing the review question, but the evidence is limited by the uncertainty in the data due to small numbers of disease positives.

An important limitation is the difficulty in understanding how the accuracy of clinical judgement for the diagnosis of dementia or cognitive impairment is related to the stage of the condition. Dementia is a progressive disease, and it is possible that GPs do not make a diagnosis of dementia until the condition has advanced to a stage where the diagnosis is relatively apparent, perhaps even to a non-medical person. This would generally lead to specificity of GP clinical judgement for dementia being higher than sensitivity. If GP clinical judgement is only accurate in advanced stages of the disease then arguably it contributes little beyond the opinion of the patients family and friends. However, attempting to analyse how the accuracy of clinical judgement varied over different stages of disease would have led to small numbers in the analysis, and would have been hampered by reporting in primary studies. The review did however include two target conditions: dementia, and cognitive impairment including dementia. The cognitive impairment target condition includes a wider spectrum of people including those with milder degrees of cognitive impairment was well as those with more advanced dementia. There was no strong evidence of difference in the sensitivity and specificity of clinical judgement for the two target conditions, with overlapping confidence intervals and the findings

limited by uncertainty, however when comparing Figure 5 and Figure 7 it is possible that the sensitivity of clinical judgement for cognitive impairment may be higher than for dementia, without loss of specificity. Since cognitive impairment is by definition inclusive of less manifest cognitive impairment than dementia, this finding may appear counter-intuitive. However, the finding that sensitivity of clinical judgement for cognitive impairment *may* be higher than the sensitivity of clinical judgement for dementia may relate to an implicit (perhaps subconscious) threshold for GPs in labelling someone as having dementia. Dementia can be a stigmatised disease and GPs may be attempting to avoid the potential psychological harm of making a false diagnosis.

An important aspect of the interpretation of the results is the threshold of clinical judgement for the diagnosis of the target conditions. It is only sensible to estimate average sensitivity and specificity at a common test threshold (Bossuyt 2013). The philosophy in this review is that the common threshold for the diagnosis of clinical judgement is the diagnostic label of dementia or cognitive impairment. A clinician makes a clinical judgement about the presence of dementia (or indeed any diagnosis) when they judge that the patient fits better, on balance, in that group of people than outside of the group, especially with regards to prognosis and response to treatment (Croft 2015). While different clinicians may formulate differing conclusions about whether or not a condition is present in a particular patient, their threshold for clinical judgement in decision making is likely to be a function of factors such as the implications of the disease (regarding prognosis or treatment), their familiarity with the patient, and the urgency of the decision. This review takes the stance that if a GP participating in one study (for instance De Lepeleire 2004) had instead been participating in a different study (for instance Wind 1995) evaluating the same patient under the same circumstances would lead to the same decision about the target condition, because the threshold for the target condition in both studies was consistently a diagnostic label of dementia. That is, that there is low intra-observer variability in the classification of a person as having dementia. Unfortunately, there is very little evidence on the intra-observer variability of the diagnosis of dementia in general practice. Available data indicate good inter-observer agreement (kappa 0.63 to 0.90) for the categorisation of dementia/no dementia (Farrer 1994; Graham 1996; Larson 1998) but these studies are based on specialists applying standardised criteria and we could not identify similar studies investigating the reliability of clinical judgement of GPs. However, the alternate view is that there is no common threshold for clinical judgement because it is a subjective test. We consider this view less plausible because this position would mean that despite identical circumstances, clinician, and patient, a different classification could be reached regarding the target condition (due to varying test threshold), and (though the play of chance is acknowledged) this is considered unlikely. If correct, the view of no common index test threshold would imply that it is not appropriate to perform bivariate meta-analysis of the data on clinical judgement.

As discussed in the previous paragraph, GPs in studies that contributed to this review may have understood their clinical judgement as equivalent to the diagnosis, with potentially important implications for their patients and so they may have been erring on the side of minimising the number of false positives, which would tend to lead to a higher specificity. From Figure 5 and Figure 7 it is plausible that studies with higher sensitivity are also those with a higher prevalence of the target condition, which may suggest that the implicit threshold for clinical judgement of dementia varies with the prevalence of the target condition. However, it is difficult to draw any firm conclusions about whether an implicit threshold for clinical judgement varies with prevalence because of limitations in the data, for example in Figure 5 two studies with similar prevalence have very different sensitivity (Pond 1994 prevalence 25% sensitivity 42%; Cooper 1992 prevalence 29% sensitivity 91%) that are both at unclear risk of bias in the flow and timing QUADAS-2 domain. In contrast, many brief cognitive assessments have the purpose of improving the identification of people with cognitive disorders and so would tend to minimise the number of false negatives, optimising sensitivity.

AUTHORS' CONCLUSIONS

Implications for practice

In practice general practitioners (GPs) are unlikely to use clinical judgement as a single test either to confirm or to exclude mild-moderate dementia or cognitive impairment, but these results indicate that clinical judgement is likely to be more specific than sensitive. When using formal cognitive assessments as part of the supplementary evaluation for people with cognitive symptoms, it may be advantageous for GPs to choose a test based on their clinical judgement. For example, if a patient has cognitive symptoms but a GP thinks they do not have dementia, clinical judgement has relatively low sensitivity and it may be useful to use a high sensitivity cognitive test to confirm that there is no cognitive disorder. On the other hand, if a GP thinks a person with cognitive symptoms does have dementia, then a high specificity test may be appropriate to help confirm the diagnosis. GP clinical judgement for cognitive impairment has higher sensitivity and lower specificity than GP clinical judgement for dementia, so a potentially useful approach in practice might be for GPs to aim to use their clinical judgement to identify possible cognitive impairment rather than dementia.

There is limited literature on the use of cognitive tests by GPs to facilitate confirming or excluding a diagnosis of dementia or cognitive impairment, but in one study, based on referrals to a memory clinic, around 26% of referrals contained an objective cognitive test (note that tests may have been done incorrectly, or been done more or less often in people who were not referred to clinic), of which around 67% of tests were the Mini-Mental State Examination (Menon 2011). In a similar study, 41% of referrals to a memory clinic indicated that a cognitive test had been done prior to referral (Wojtowicz 2015). Unfortunately there is currently very limited evidence on the diagnostic accuracy of different tests in a general practice population, which is important because test accuracy is related to prevalence of disease. However, one pragmatic approach would be for GPs to use a test that they are familiar with, such as the Mini-Mental State Examination (Creavin 2016), at a lower test threshold (indicating more severe disease) to confirm objective cognitive impairment in a person who a GP thinks has dementia, and a higher threshold to confirm the absence of objective cognitive impairment in a person who a GP thinks does not have dementia. One limitation of using objective tests with high sensitivity to mitigate against the relatively poor sensitivity of GP clinical judgement, especially for dementia, is that it would increase the number of false positives, people who fail the test without having dementia, who would then require further tests or a referral to confirm whether they do, or do not, have dementia.

However, on the other hand, many people with cognitive problems who seek help, find it helpful to have an explanation for those problems, and an explanation can provide access to support and interventions.

Dementia is an uncommon disease in general practice and it has been projected that most GPs could expect one to two new cases each year, per physician (Iliffe 2009). If a GP thinks a patient has dementia then there may be merit in using a brief cognitive test to measure objective cognitive impairment, or to objectively measure and quantify progression in cognitive decline, because the probability of dementia based on GP clinical judgement alone is not high enough to confirm dementia. It is also important to objectively quantify cognitive functioning in someone with symptoms of dementia who is judged by their GP to be disease-free, because the sensitivity of clinical judgement is too low to definitively exclude cognitive problems. Cognitive impairment has a variety of aetiologies, including dementia, as described in Target condition being diagnosed, so further work-up to clarify the aetiology of the problem may be required in someone who a GP thinks has clinically significant symptoms.

A final implication for practice relates particularly to the very oldest people with dementia who often have multimorbidity. Some people with cognitive symptoms in the late stages of their life, may not want to be troubled by tests and investigations, and may be unwilling to undergo formal cognitive testing, or find the process demeaning and belittling. For these people, a positive clinical judgement about the presence of dementia may be sufficient, and may be helpful to frame decision making and planning about the rest of their care in a holistic way.

Implications for research

An important unanswered question from this review is what is the accuracy of GP clinical judgement for the diagnosis of dementia in people with symptoms compared to those without. Most studies in this review did not distinguish between people attending their GP with symptoms of dementia and those attending their GP for other reasons. This is especially important because the prevalence of dementia in people attending their GP with symptoms may be higher than the 30% calculated in this review, as well as also altering the spectrum of disease in people who do have dementia (being more severe in people who are presenting with symptoms). A phenomenon that we have observed in clinical practice over time and especially in recent years, is people presenting to healthcare providers with less severe symptoms than would have been the case in years gone by (Bell 2015; Menon 2011; Wojtowicz 2015). A consequence of this is that the accuracy of clinical judgement may change in the future as the prevalence of underlying disease changes, and GPs may become more reliant on the use of objective cognitive tests. Contemporary research to quantify the prevalence of dementia and mild cognitive impairment in people presenting to their GP with symptoms would be useful. Overall it is possible the estimates of diagnostic accuracy in this review are likely an underestimate of diagnostic accuracy when clinical judgement is applied only in people with symptoms.

There are at least two further important unanswered questions from this review. Firstly, what is the comparative accuracy of clinical judgement compared to brief cognitive tests for diagnosing dementia? Direct comparison of test accuracy in a single study is preferable to indirect comparisons but requires additional methodological considerations to ensure robust study design. Secondly, what is the evidence for heterogeneity in the aspects of study design that we investigated in this review? To investigate heterogeneity in test accuracy further, future studies could attempt to apply more than one definition, could elicit retrospective judgement in a randomly selected sample of people who had already had prospective GP clinical judgement and who were intended to receive the reference standard assessment (but had not yet), could blind some (randomly assigned) participating GPs to using the medical records for the retrospective judgement and others to not, and could take steps to attempt full verification of the reference standard and provide extended follow-up of the medical records for those people who were not verified. Database studies using routinely collected data may (with limitations) be a particularly useful methodological development to overcome the challenges of extended follow-up over long periods of time, and should be considered for future test accuracy studies in this area.

ACKNOWLEDGEMENTS

We would like to thank consumer reviewer Kit Byatt, clinical peer reviewer Dimity Pond and one other clinical peer revIewer who wishes to remain anonymous for their comments and feedback. We also thank copy editor Luisa M Fernandez Mauleffinch from Cochrane Copy Edit Support.

REFERENCES

References to studies included in this review

Brayne 1990 *{published data only}*

Brayne C, Calloway P. An epidemiological study of dementia in a rural population of elderly women. *British Journal of Psychiatry* 1989;**155**:214-9. [PMID: 2597917]

* Brayne C, Calloway P. The case identification of dementia in the community: a comparison of methods. *International Journal of Geriatric Psychiatry* 1990;**5**:309-16. [DOI: 10.1002/gps.930050507]

Cooper 1992 *{published data only}*

Cooper B, Bickel H, Schäufele M. Dementia diseases and minor cognitive impairments in elderly patients in general practice. Results of a cross-sectional study [Demenzerkrankungen und leichtere kognitive Beeinträchtigungen bei älteren Patienten in der ärztlichen Allgemeinpraxis Ergebnisse einer Querschnittsuntersuchung]. *Nervenarzt* 1992;**63**(9):551-60. [PMID: 1407226]

* Cooper B, Bickel H, Schäufele M. The ability of general practitioners to detect dementia and cognitive impairment in their elderly patients: a study in Mannheim. *International Journal of Geriatric Psychiatry* 1992;**7**:591-8. [DOI: 10.1002/gps.930070809]

Creavin 2021 *{published data only}***doi.org/10.1101/2020.11.20.20234062**

Creavin ST, Haworth J, Fish M, Cullum S, Bayer A, Purdy S, et al. Clinical judgment of GPs for the diagnosis of dementia: a diagnostic test accuracy study. *BJGP Open* 2021;**5**(5):BJGPO.2021.0058. [PMID: 34315715]

De Lepeleire 2004 *{published data only}*

De Lepeleire J, Aertgeerts B, Umbach I, Pattyn P, Tamsin F, Nestor L, et al. The diagnostic value of IADL evaluation in the detection of dementia in general practice. *Aging & Mental Health* 2004;**8**(1):52-7. [PMID: 14690868]

Eefsting 1996 *{published data only}*

Eefsting JA, Boersma F, Van den Brink W, Van Tilburg W. Differences in prevalence of dementia based on community survey and general practitioner recognition. *Psychological Medicine* 1996;**26**(6):1223-3. [PMID: 8931168]

O'Connor 1988 *{published data only}*

* O'Connor DW, Pollitt PA, Hyde JB, Brook CP, Reiss BB, Roth M. Do general practitioners miss dementia in elderly patients? *BMJ* 1988;**297**(6656):1107-10. [PMID: 3143447]

O'Connor DW, Pollitt PA, Hyde JB, Fellows JL, Miller ND, Brook CP, et al. The prevalence of dementia as measured by the Cambridge Mental Disorders of the Elderly Examination. *Acta Psychiatrica Scandinavica* 1989;**79**(2):190-8. [PMID: 2923012]

Pentzek 2009 *{published data only}*

Kaduszkiewicz H, Zimmermann T, Van den Bussche H, Bachmann C, Wiese B, Bickel H, et al. Do general practitioners recognize mild cognitive impairment in their patients? *Journal of Nutrition, Health & Aging* 2010;**14**(8):697-702. [DOI: 20922348]

Pentzek M, Fuchs A, Wiese B, Abholz HH. Which information do GPs use to rate the cognitive status of elderly non-demented patients? [Welche Informationen nutzen Hausärzte zur Einschätzung des kognitiven Status älterer nicht dementer Patienten?]. *Psychiatrische Praxis* 2010;**37**(8):377-83. [PMID: 20703983]

* Pentzek M, Wollny A, Wiese B, Jessen F, Haller F, Maier W, et al. Apart from nihilism and stigma: what influences general practitioners' accuracy in identifying incident dementia? *American Journal of Geriatric Psychiatry* 2009;**17**(11):965-75. [PMID: 20104054]

Pond 1994 *{published data only}*

Pond CD, Mant A, Kehoe L, Hewitt H, Brodaty H. General practitioner diagnosis of depression and dementia in the elderly: can academic detailing make a difference? *Family Practice* 1994;**11**(2):141-7. [PMID: 7958576]

Rondeau 2008 *{published data only}*

Rondeau V, Allain H, Bakchine S, Bonet P, Brudon F, Chauplannaz G, et al. General practice-based intervention for suspecting and detecting dementia in France: a cluster randomized controlled trial. *Dementia* 2008;**7**(4):433–50. [DOI: 10.1177/1471301208096628]

Valcour 2000 *{published data only (unpublished sought but not used)}*

Valcour VG, Masaki KH, Curb JD, Blanchette PL. The detection of dementia in the primary care setting. *Archives of Internal Medicine* 2000;**160**(19):2964-8. [PMID: 11041904]

Wind 1995 *{published data only}*

* Wind AW, Schellevis FG, van Staveren G, Scholten RJ, Hooijer C, Jonker C, et al. Determinants of the judgement of general practitioners on dementia. *International Journal of Geriatric Psychiatry* 1995;**10**:767-76. [DOI: 10.1002/gps.930100907]

Wind AW, van Staveren G, Schellevis FG, Jonker C, Eijk JT. The validity of the judgement of general practitioners on dementia. *International Journal of Geriatric Psychiatry* 1994;**9**:543-9. [DOI: 10.1002/gps.930090707]

References to studies excluded from this review

Aldus 2018 *{published data only}*

Aldus C, Arthur A, Fox C, Brayne C, Matthews F, Dening T, et al. Cognitive function and ageing study II dementia diagnosis study (CADDY): the prevalence, causes and consequences of dementia undetected or undiagnosed in primary care in England. *Alzheimer's & Dementia* 2018;**14**:573–4.

Belmin 2012 *{published data only}*

Belmin J, Min L, Roth C, Reuben D, Wenger N. Assessment and management of patients with cognitive impairment and

dementia in primary care. *Journal of Nutrition, Health & Aging* 2012;**16**:462–7.

Borson 2006 *{published data only}*

Borson S, Scanlan JM, Watanabe J, Tu SP, Lessig M. Improving identification of cognitive impairment in primary care. *International Journal of Geriatric Psychiatry* 2006;**21**:349–55.

Bushnell 2004 *{published data only}*

Bushnell J, MaGPIe Research Group. Frequency of consultations and general practitioner recognition of psychological symptoms. *British Journal of General Practice* 2004;**54**:838–43.

Camicioli 2000 *{published data only}*

Camicioli R, Willert P, Lear J, Grossmann S, Kaye J, Butterfield P. Dementia in rural primary care practices in Lake County, Oregon. *Journal of Geriatric Psychiatry and Neurology* 2000;**13**:87-92.

Chong 2016 *{published data only}*

Chong SA, Abdin E, Vaingankar J, Ng LL, Subramaniam M. Diagnosis of dementia by medical practitioners: a national study among older adults in Singapore. *Aging & Mental Health* 2016;**20**:1271–6.

De Lepeleire 2005 *{published data only}*

De Lepeleire J, Heyrman J, Baro F, Buntinx F. A combination of tests for the diagnosis of dementia had a significant diagnostic value. *Journal of Clinical Epidemiology* 2005;**58**:217–25.

Dilts 2003 *{published data only}*

Dilts SL, Mann N, Dilts JG. Accuracy of referring psychiatric diagnosis on a consultation-liaison service. *Psychosomatics* 2003;**44**:407–11.

Dinesen 1997 *{published data only}*

Dinesen O, Frijs-Madsen B, Almbjerg F, Fromholt P, Torpdahl P. Dementia diagnosis in general practice [Demensdiagnostik i almen praksis]. *Ugeskrift for Læger* 1997;**159**:5795–9.

Engedal 1989 *{published data only}*

Engedal K, Gilje K, Lilleaas F. Diagnostic evaluation of the mentally impaired elderly living at home. *Scandinavian Journal of Primary Health Care* 1989;**7**:5–11.

Fichter 1995 *{published data only}*

Fichter MM, Meller I, Schröppel H, Steinkirchner R. Dementia and cognitive impairment in the oldest old in the community. Prevalence and comorbidity. *British Journal of Psychiatry* 1995;**166**:621–9.

Hara 2013 *{published data only}*

Hara J, Macias D, Russell J, Fortier D, Holnagel D, Keeble C, et al. Prevalence of cognitive impairment based on the annual wellness visit. *Alzheimer's & Dementia* 2013;**9**:119.

Hessler 2014 *{published data only}*

Hessler J, Bronner M, Etgen T, Ander K, Foerstl H, Poppert H, et al. Suitability of the 6CIT as a screening test for dementia in primary care patients. *Aging & Mental Health* 2014;**18**:515–20.

Hopman-Rock 2001 *{published data only}*

Hopman-Rock M, Tak ECPM, Staats PGM. Development and validation of the Observation List for early signs of Dementia (OLD). *International Journal of Geriatric Psychiatry* 2001;**16**:406–14.

Jacinto 2009 *{published data only}*

Jacinto AF, Nitrini R, Brucki SMD, Porto CS. Detection of cognitive impairment in the elderly by general practitioners in Brazil. *Alzheimer's & Dementia* 2009;**5**:189–90.

Jansen 2007 *{published data only}*

Jansen APD, Hout HPJ van, Nijpels G, Marwijk HWJ van, Vet HCW de, Stalman WAB. Yield of a new method to detect cognitive impairment in general practice. *International Journal of Geriatric Psychiatry* 2007;**22**:590–7.

Juva 1994 *{published data only}*

Juva K, Sulkava R, Erkinjuntti T, Ylikoski R, Valvanne J, Tilvis R. Staging the severity of dementia: comparison of clinical (CDR, DSM-III-R), functional (ADL, IADL) and cognitive (MMSE) scales. *Acta Neurologica Scandinavica* 1994;**90**(4):293-8. [PMID: 7839817]

Kurz 1999 *{published data only}*

Kurz X, Broers M, Scuvee-Moreau J, Salmon E, Ventura M, Pepin JL, et al. Methodological issues in a cost-of-dementia study in Belgium: the NAtional Dementia Economic Study (NADES). *Acta Neurologica Belgica* 1999;**99**:167–75.

Leung 2007 *{published data only}*

Leung WC. GPs' diagnosis of dementia. *British Journal of General Practice* 2000;**50**:666.

Lionis 2001 *{published data only}*

Lionis C, Tzagournissakis M, Iatraki E, Kozyraki M, Antonakis N, Plaitakis A. Are primary care physicians able to assess dementia? An estimation of their capacity after a short-term training program in rural Crete. *American Journal of Geriatric Psychiatry* 2001;**9**:315.

Livingston 1990 *{published data only}*

Livingston G, Sax K, Willison J, Blizard B, Mann A. The Gospel Oak Study stage II: the diagnosis of dementia in the community. *Psychological Medicine* 1990;**20**:881–91.

Löppönen 2003 *{published data only}*

Löppönen M, Räihä I, Isoaho R, Vahlberg T, Kivelä SL. Diagnosing cognitive impairment and dementia in primary health care - a more active approach is needed. *Age and Ageing* 2003;**32**(6):606-12. [PMID: 14600001]

Mant 1988 *{published data only}*

Mant A, Eyland EA, Pond DC, Saunders NA, Chancellor AH. Recognition of dementia in general practice: comparison of general practitioners' opinions with assessments using the mini-mental state examination and the Blessed dementia rating scale. *Family Practice* 1988;**5**:184–8.

Mok 2004 *{published data only}*

Mok W, Chow TW, Zheng L, Mack WJ, Miller C. Clinicopathological concordance of dementia diagnoses by community versus tertiary care clinicians. *American Journal of Alzheimer's Disease and Other Dementias* 2004;**19**(3):161-5. [DOI: 10.1177/153331750401900309]

Noda 2018 *{published data only}*

Noda H, Yamagishi K, Ikeda A, Asada T, Iso H. Identification of dementia using standard clinical assessments by primary care physicians in Japan. *Geriatrics & Gerontology International* 2018;**18**:738–44.

Olafsdóttir 2000 *{published data only}*

Olafsdóttir M, Skoog I, Marcusson J. Detection of dementia in primary care: the Linköping study. *Dementia and Geriatric Cognitive Disorders* 2000;**11**(4):223-9. [PMID: 10867449]

Pittmann 1992 *{published data only}*

Pittman J, Andrews H, Tatemichi T, Link B, Struening E, Stern Y, et al. Diagnosis of dementia in a heterogeneous population. A comparison of paradigm-based diagnosis and physician's diagnosis. *Archives of Neurology* 1992;**49**:461–7.

Schaub 2003 *{published data only}*

Schaub RT, Linden M, Copeland JRM. A comparison of GMS-A/AGECAT, DSM-III-R for dementia and depression, including sub-threshold depression (SD)–results from the Berlin Aging Study (BASE). *International Journal of Geriatric Psychiatry* 2003;**18**:109–17.

Tierney 2014 *{published data only}*

Tierney MC, Naglie G, Upshur R, Jaakkimainen L, Moineddin R, Charles J, et al. Factors associated with primary care physicians' recognition of cognitive impairment in their older patients. *Alzheimer Disease and Associated Disorders* 2014;**28**:320–5.

van Hout 1999 *{published data only}*

van Hout H, Vernooij-Dassen M, Hoefnagels W, Grol R. Use of mini-mental state examination by GPs to diagnose dementia may be unnecessary. *BMJ* 1999;**319**:190.

van Hout 2000 *{published data only}*

van Hout H, Vernooij-Dassen M, Poels P, Hoefnagels WH, Grol RP. Are general practitioners able to accurately diagnose dementia and identify Alzheimer's disease? A comparison with an outpatient memory clinic. *British Journal of General Practice* 2000;**50**:311–2.

van Hout 2001 *{published data only}*

van Hout H. Applicability of diagnostic recommendations on dementia in family practice. *International Journal for Quality in Health Care* 2001;**13**:127–33.

van Hout 2002 *{published data only}*

van Hout HP, Vernooij-Dassen MJ, Hoefnagels WH, Kuin Y, Stalman WA, Moons KG, et al. Dementia: predictors of diagnostic accuracy and the contribution of diagnostic recommendations. *Journal of Family Practice* 2002;**51**:693–9.

van Hout 2003 *{published data only}*

van Hout HPJ, Vernooij-Dassen MJ, Hoefnagels WH, Kuin Y, Stalman WA, Moons KGM, et al. The diagnostic value of the recommendations from the NHG Standard for Dementia [De diagnostische waarde van de aanbevelingen uit de NHG-Standaard Dementie]. *Huisarts en Wetenschap* 2003;**46**:71–8.

van Hout 2006 *{published data only}*

van Hout HP, Vernooij-Dassen MJ, Jansen DA, Stalman WA. Do general practitioners disclose correct information to their patients suspected of dementia and their caregivers? A prospective observational study. *Aging & Mental Health* 2006;**10**:151–5.

van Hout 2007a *{published data only}*

van Hout HP, Vernooij-Dassen MJ, Stalman WA. Diagnosing dementia with confidence by GPs. *Family Practice* 2007;**24**:616–21.

van Hout 2007b *{published data only}*

van Hout H, Vernooij-Dassen M, Jansen D, Stalman W. Do GPs provide correct information to patients suffering from dementia and their carers? [Geven huisartsen aan dementerende patiënten en hun verzorgers de juiste informatie?]. *Huisarts en Wetenschap* 2007;**50**:424–30.

Waldorff 2005 *{published data only}*

Waldorff FB, Rishøj S, Waldemar G. Identification and diagnostic evaluation of possible dementia in general practice. *Scandinavian Journal of Primary Health Care* 2005;**23**:221–6.

Wang 2017 *{published data only}*

Wang SZ, Tan GC, Wang XF, De Roza JG, Lim LL, Kandiah N. Experience with a community-based multidisciplinary memory clinic: a primary care perspective. *Annals of the Academy of Medicine, Singapore* 2017;**46**:321–3.

Wilkins 2007 *{published data only}*

Wilkins CH, Wilkins KL, Meisel M, Depke M, Williams J, Edwards DF. Dementia undiagnosed in poor older adults with functional impairment. *Journal of the American Geriatrics Society* 2007;**55**:1771–6.

Additional references

Almond 2009

Almond SC, Summerton N. Diagnosis in general practice. Test of time. *BMJ* 2009;**15**(338):b1878. [PMID: 19528115]

American Psychiatric Association 2000

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders Fourth TR. Arlington (VA): American Psychiatric Association, 2000.

Barberger-Gateau 1999

Barberger-Gateau P, Fabrigoule C, Helmer C, Rouch I, Dartigues JF. Functional impairment in instrumental activities of daily living: an early clinical sign of dementia? *Journal of the American Geriatrics Society* 1999;**47**:456–62.

Barraclough 2006

Barraclough K. Medical intuition. *BMJ* 2006;**332**:497.

Bell 2015

Bell S, Harkness K, Dickson JM, Blackburn D. A diagnosis for £55: what is the cost of government initiatives in dementia case finding. *Age and Ageing* 2015;**44**(2):344-5.

Blaeuer 2013

Blaeuer SR, Bally K, Tschudi P, Martina B, Zeller A. Acute cough illness in general practice - predictive value of clinical judgement and accuracy of requesting chest x-rays. *Praxis* 2013;**102**(21):1287-92.

Blessed 1968

Blessed G, Tomlinson BE, Roth M. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry* 1968;**114**:797–811.

Body 2014

Body R, Cook G, Burrows G, Carley S, Lewis PS. Can emergency physicians "rule in" and "rule out" acute myocardial infarction with clinical judgement? *Emergency Medicine Journal* 2014;**31**(11):872-6.

Boise 2004

Boise L, Neal MB, Kaye J. Dementia assessment in primary care: results from a study in three managed care systems. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 2004;**59**:M621–6.

Bossuyt 2013

Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11: Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C, editor(s). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9. The Cochrane Collaboration, 2013. Available from srdta.cochrane.org/.

Boustani 2005

Boustani M, Callahan CM, Unverzagt FW, Austrom MG, Perkins AJ, Fultz BA, et al. Implementing a screening and diagnosis program for dementia in primary care. *Journal of General Internal Medicine* 2005;**20**:572–7.

Bowers 1990

Bowers J, Jorm AF, Henderson S, Harris P. General practitioners' detection of depression and dementia in elderly patients. *Medical Journal of Australia* 1990;**153**:192–6.

Brayne 2007

Brayne C, Fox C, Boustani M. Dementia screening in primary care: is it time? *JAMA* 2007;**298**(20):2409-11.

Brayne 2012

Brayne C, Davis D. Making Alzheimer's and dementia research fit for populations. *Lancet* 2012;**380**(9851):1441-3.

Brodaty 2002

Brodaty H, Pond D, Kemp NM, Luscombe G, Harding L, Berman K, et al. The GPCOG: a new screening test for dementia designed for general practice. *Journal of the American Geriatrics Society* 2002;**50**(3):530-4.

Brooke 1999

Brooke P, Bullock R. Validation of a 6 item cognitive impairment test with a view to primary care usage. *International Journal of Geriatric Psychiatry* 1999;**14**:936–40.

Brown 2009

Brown J, Pengas G, Dawson K, Brown LA, Clatworthy P. Self administered cognitive screening test (TYM) for detection of Alzheimer's disease: cross sectional study. *BMJ* 2009;**338**:b2030.

Brush 2017

Brush JE, Sherbino J, Norman GR. How expert clinicians intuitively recognize a medical diagnosis. *American Journal of Medicine* 2017;**130**:629–34.

Burns 2013

Burns A. Alistair Burns and 51 colleagues reply to David Le Couteur and colleagues. *BMJ* 2013;**347**:f6125.

Callahan 1995

Callahan CM, Hendrie HC, Tierney WM. Documentation and evaluation of cognitive impairment in elderly primary care patients. *Annals of Internal Medicine* 1995;**122**:422–9.

Charlin 2000

Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Academic Medicine* 2000;**75**(1):182-90.

Chen 2014

Chen L, Reed C, Happich M, Nyhuis A, Lenox-Smith A. Health care resource utilisation in primary care prior to and after a diagnosis of Alzheimer's disease: a retrospective, matched case-control study in the United Kingdom. *BMC Geriatrics* 2014;**14**:76.

Chodosh 2004

Chodosh J, Petitti DB, Elliott M, Hays RD, Crooks VC, Reuben DB, et al. Physician recognition of cognitive impairment: evaluating the need for improvement. *Journal of the American Geriatrics Society* 2004;**52**:1051–9.

Chu 2006

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006;**59**(12):1331-2; author reply 1332-3.

Copeland 1986

Copeland JR, Dewey ME, Griffiths-Jones HM. A computerized psychiatric diagnostic system and case nomenclature for elderly subjects: GMS and AGECAT. *Psychological Medicine* 1986;**16**(1):89-99.

Covidence [Computer program]

Veritas Health Innovation Covidence. Version accessed prior to 25 May 2022. Melbourne, Australia: Veritas Health Innovation, 2022. Available at 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]

Croft 2015

Croft P, Altman DG, Deeks JJ, Dunn KM, Hay AD, Hemingway H, et al. The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice. *BMC Medicine* 2015;**13**:20.

Di 2013

Di Somma S, Magrini L, De Berardinis B, Marino R, Ferri E, Moscatelli P, et al. Additive value of blood neutrophil gelatinase-associated lipocalin to clinical judgement in acute kidney injury diagnosis and mortality prediction in patients hospitalized from the emergency department. *Critical Care* 2013;**17**(1):R29.

Elstein 2009

Elstein AS. Thinking about diagnostic thinking: a 30-year perspective. *Advances in Health Sciences Education* 2009;**14**(1 Suppl):7-18.

Ensor 2018

Ensor J, Deeks JJ, Martin EC, Riley RD. Meta-analysis of test accuracy studies using imputation for partial reporting of multiple thresholds. *Research Synthesis Methods* 2018;**9**:100–15.

Erkinjuntti 1997

Erkinjuntti T, Østbye T, Steenhuis R, Hachinski V. The effect of different diagnostic criteria on the prevalence of dementia. *New England Journal of Medicine* 1997;**337**:1667–74.

Farrer 1994

Farrer LA, Cupples LA, Blackburn S, Kiely DK, Auerbach S, Growdon JH, et al. Interrater agreement for diagnosis of Alzheimer's disease: the MIRAGE study. *Neurology* 1994;**44**:652–6.

Folstein 1975

Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975;**12**:189–98.

Ford 2018

Ford E, Greenslade N, Paudyal P, Bremner S, Smith HE, Banerjee S, et al. Predicting dementia from primary care records: a systematic review and meta-analysis. *PLOS One* 2018;**13**(3):e0194735.

Fox 2013

Fox C, Lafortune L, Boustani M, Brayne C. The pros and cons of early diagnosis in dementia. *British Journal of General Practice* 2013;**63**(612):e510-2.

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**:81.

Ganguli 2004

Ganguli M, Rodriguez E, Mulsant B, Richards S, Pandav R, Bilt JV, et al. Detection and management of cognitive impairment in primary care: the Steel Valley Seniors Survey. *Journal of the American Geriatrics Society* 2004;**52**:1668–75.

Gauthier 2006

Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. *Lancet* 2006;**367**(9518):1262-70.

Gill 2014

Gill PJ, Roberts NW, Wang KY, Heneghan C. Development of a search filter for identifying studies completed in primary care. *Family Practice* 2014;**31**(6):739-45.

Graham 1996

Graham JE, Rockwood K, Beattie BL, McDowell I, Eastwood R, Gauthier S. Standardization of the diagnosis of dementia in the Canadian Study of Health and Aging. *Neuroepidemiology* 1996;**15**:246–56.

Harbord 2007

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;**8**(2):239-51.

Harskamp 2019

Harskamp RE, Laeven SC, Himmelreich JC, Lucassen WAM, van Weert HC. Chest pain in general practice: a systematic review of prediction rules. *BMJ Open* 2019;**9**(2):e027081.

Heneghan 2009

Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. *BMJ* 2009;**338**:b946.

Heyrman 1990

Heyrman J, Dessers L, Munter MB, de Haepers K, Craenen J. Functional Status Assessment in the Elderly. In: WONCA Classification Committee, editors(s). Functional Status Measurement in Primary Care. New York: Springer, 1990:213–21.

Hughes 1982

Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. *British Journal of Psychiatry* 1982;**140**:566-72.

ICD 1993

World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research. Geneva: World Health Organization, 1993.

Iliffe 1990

Iliffe S, Booroff A, Gallivan S, Goldenberg E, Morgan P, Haines A. Screening for cognitive impairment in the elderly using the mini-mental state examination. *British Journal of General Practice* 1990;**40**:277–9.

Iliffe 2009

Iliffe S, Robinson L, Brayne C, Goodman C, Rait G, Manthorpe J, et al. Primary care and dementia: 1. diagnosis, screening and disclosure. *International Journal of Geriatric Psychiatry* 2009;**24**:895–901.

Iliffe 2014

Iliffe S. General practitioners should be conducting targeted screening for dementia in people aged 65 to 74. *Journal of Primary Health Care* 2014;**6**(3):247-9.

Jones 2010

Jones R, Barraclough K, Dowrick C. When no diagnostic label is applied. *BMJ* 2010;**340**:c2683.

Kaduszkiewicz 2014

Kaduszkiewicz H, Eisele M, Wiese B, Prokein J, Luppa M, Luck T, et al. Prognosis of mild cognitive impairment in general practice: results of the German AgeCoDe study. *Annals of Family Medicine* 2014;**12**(2):158-65.

Kahneman 2011

Kahneman D. Thinking, Fast and Slow. 1st edition. New York: Farrar, Straus and Giroux, 2011.

Kawas 1994

Kawas C, Segal J, Stewart WF, Corrada M, Thal LJ. A validation study of the Dementia Questionnaire. *Archives of Neurology* 1994;**51**:901–6.

Kawas 2015

Kawas CH, Kim RC, Sonnen J, Bullain SS, Trieu T. Multiple pathologies are common and related to dementia in the oldest-old. *Neurology* 2015;**85**(6):535-42.

Koch 2010

Koch T, Iliffe S. Rapid appraisal of barriers to the diagnosis and management of patients with dementia in primary care: a systematic review. *BMC Family Practice* 2010;**11**:52.

Lambe 2016

Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety* 2016;**25**:808–20.

Larson 1998

Larson EB, McCurry SM, Graves AB, Bowen JD, Rice MM, McCormick WC, et al. Standardization of the clinical diagnosis of the dementia syndrome and its subtypes in a cross-national study: the Ni-Hon-Sea experience. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 1998;**53**(4):M313–9.

Le 2013

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.

Lehman 2015

Lehman R. Siddharta Mukherjee's three laws of medicine. *BMJ* 2015;**351**:h6708.

Macaskill 2010

Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editor(s). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from srdta.cochrane.org.

Magaziner 2000

Magaziner J, German P, Zimmerman SI, Hebel JR, Burton L, Gruber-Baldini AL, et al. The prevalence of dementia in a statewide sample of new nursing home admissions aged 65 and older: diagnosis by expert panel. Epidemiology of Dementia in Nursing Homes Research Group. *Gerontologist* 2000;**40**(6):663-72.

Matthews 2013

Matthews FE, Arthur A, Barnes LE, Bond J, Jagger C, Robinson L, et al. A two-decade comparison of prevalence of dementia in individuals aged 65 years and older from three geographical areas of England: results of the Cognitive Function and Ageing Study I and II. *Lancet* 2013;**382**(9902):1405-12.

McKeith 1996

McKeith IG, Galasko D, Kosaka K, Perry EK, Dickson D W, Hansen LA, et al. Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the consortium on DLB international workshop. *Neurology* 1996;**47**(5):1113-24.

McKeith 2005

McKeith IG, Dickson DW, Lowe J, Emre M, O'Brien JT, Feldman H, et al. Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology* 2005;**65**(12):1863-72.

McKhann 1984

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984;**34**(7):939-44.

McKhann 2011

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, 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. *Alzheimer's & Dementia* 2011;**7**(3):263-9.

Menon 2011

Menon R, Larner AJ. Use of cognitive screening instruments in primary care: the impact of national dementia directives (NICE/SCIE, National Dementia Strategy). *Family Practice* 2011;**28**(3):272–6.

Mitchell 2011

Mitchell AJ, Meader N, Pentzek M. Clinical recognition of dementia and cognitive impairment in primary care: a meta-analysis of physician accuracy. *Acta Psychiatrica Scandinavica* 2011;**124**(3):165-83.

Morris 1993

Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 1993;**43**:2412–4.

Nasreddine 2005

Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 2005;**53**(4):695-9.

Neary 1994

Neary D, Brun A, Englund B, Gustafson L, Passant U, Mann DMA, et al. Clinical and neuropathological criteria for frontotemporal dementia. The Lund and Manchester Groups. *Journal of Neurology, Neurosurgery, and Psychiatry* 1994;**57**(4):416-8.

Neuropathology 2001

Neuropathology Group of the Medical Research Council Cognitive Function and Ageing Study (MRC CFAS). Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. *Lancet* 2001;**357**(9251):169-75.

Norman 2007

Norman G, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. *Medical Education* 2007;**41**(12):1140-5.

O'Connor 1993

O'Connor DW, Fertig A, Grande MJ, Hyde JB, Perry JR, Roland MO, et al. Dementia in general practice: the practical consequences of a more positive approach to diagnosis. *Journal of the Royal College of General Practitioners* 1993;**43**(370):185-8.

Patel 2020

Patel A, Cooper NJ, Freeman SC, Sutton AJ. Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data. *Research Synthesis Methods* 2020;**12**(1):34-44. [DOI: https://doi.org/10.1002/jrsm.1439]

Pentzek 2009

Pentzek M, Fuchs A, Wiese B, Cvetanovska-Pllashniku G, Haller F, Maier W, et al. General practitioners' judgment of their elderly patients' cognitive status. *Journal of General Internal Medicine* 2009;**24**(12):1314-7. [PMID: 19844763]

Pentzek 2019

Pentzek M, Wagner M, Abholz HH, Bickel H, Kaduszkiewicz H, Wiese B, et al. The value of the GP's clinical judgement in predicting dementia: a multicentre prospective cohort study among patients in general practice. *British Journal of General Practice* 2019;**69**(688):e786-93.

Petersen 1999

Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology* 1999;**56**(3):303-8.

Petersen 2004

Petersen RC. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine* 2004;**256**(3):183-94.

Pfeiffer 1975

Pfeiffer E. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society* 1975;**23**:433–41.

Pond 2013

Pond CD, Mate KE, Phillips J, Stocks NP, Magin PJ, Weaver N, et al. Predictors of agreement between general practitioner detection of dementia and the revised Cambridge Cognitive Assessment (CAMCOG-R). *International Psychogeriatrics* 2013;**25**(10):1639-47.

Ramakers 2007

Ramakers IH, Visser PJ, Aalten P, Boesten JH, Metsemakers JF, Jolles J, et al. Symptoms of preclinical dementia in general practice up to five years before dementia diagnosis. *Dementia and Geriatric Cognitive Disorders* 2007;**24**:300–6.

Rasmussen 2013

Rasmussen J. Improving diagnosis and management of dementia in primary care. *Progress in Neurology and Psychiatry* 2013;**17**(6):4-6.

Rasmussen 2014

Rasmussen J. General practitioners should be conducting targeted screening for dementia in people aged 65 to 74: yes. *Journal of Primary Health Care* 2014;**6**(3):245-7.

Reitsma 2005

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005;**58**(10):982-90.

Review Manager 2020 [Computer program]

The Cochrane Collaboration Review Manager 5 (RevMan 5). Version 5.4. The Cochrane Collaboration, 2020.

Román 1993

Román GC, Tatemichi TK, Erkinjuntti T, Cummings JL, Masdeu JC, Garcia JH, et al. Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop. *Neurology* 1993;**43**(2):250-60.

Roth 1986

Roth M, Tym E, Mountjoy CQ, Huppert FA, Hendrie H, Verma S, et al. CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *British Journal of Psychiatry* 1986;**149**:698-709.

Russell 2013

Russell P, Banerjee S, Watt J, Adleman R, Agoe B, Burnie N, et al. Improving the identification of people with dementia in primary care: evaluation of the impact of primary care dementia coding guidance on identified prevalence. *BMJ Open* 2013;**3**(12):e004023.

Savva 2009

Savva GM, Wharton SB, Ince PG, Forster G, Matthews FE, Brayne C. Age, neuropathology, and dementia. *New England Journal of Medicine* 2009;**360**(22):2302-9.

Sibbett 2017

Sibbett RA, Russ TC, Deary IJ, Starr JM. Dementia ascertainment using existing data in UK longitudinal and cohort studies: a systematic review of methodology. *BMC Psychiatry* 2017;**17**:239.

Stata [Computer program]

Stata. Version 15. College Station, TX, USA: StataCorp, 2017. Available at www.stata.com.

Takwoingi 2015

Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. *Evidence Based Mental Health* 2015;**18**:103-9.

van Blijswijk 2018

van Blijswijk SC, Blom JW, de Craen AJ, den Elzen WP, Gussekloo J. Prediction of functional decline in community-dwelling older persons in general practice: a cohort study. *BMC Geriatrics* 2018;**18**(1):140.

van den Dungen 2012

van den Dungen P, van Marwijk HW, van der Horst HE, Moll van Charante EP, Macneil Vroomen J, van de Ven PM, et al. The accuracy of family physicians' dementia diagnoses at different stages of dementia: a systematic review. *International Journal of Geriatric Psychiatry* 2012;**27**(4):342-54.

Waldorff 2012

Waldorff FB, Siersma V, Vogel A, Waldemar G. Subjective memory complaints in general practice predicts future dementia: a 4-year follow-up study. *International Journal of Geriatric Psychiatry* 2012;**27**(11):1180-8.

Whiting 2011a

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. *Annals of Internal Medicine* 2011;**155**(8):529-36.

Whiting 2011b

Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. *Journal of Clinical Epidemiology* 2011;**64**(6):602-7.

Winblad 2004

Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund L-O, et al. Mild cognitive impairment - beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine* 2004;**256**(3):240-6.

Wojtowicz 2015

Wojtowicz A, Larner AJ. General Practitioner Assessment of Cognition: use in primary care prior to memory clinic referral. *Neurodegenerative Disease Management* 2015;**5**(6):505-10.

Woolley 2013

Woolley A, Kostopoulou O. Clinical intuition in family medicine: more than first impressions. *Annals of Family Medicine* 2013;**11**(1):60-6.

Ydstebø 2015

Ydstebø AE, Bergh S, Selbæk G, Benth JŠ, Lurås H, Vossius C. The impact of dementia on the use of general practitioners among the elderly in Norway. *Scandinavian Journal of Primary Health Care* 2015;**33**:199–205.

Zaudig 1991

Zaudig M, Mittelhammer J, Hiller W, Pauls A, Thora C, Morinigo A, et al. SIDAM–A structured interview for the diagnosis of dementia of the Alzheimer type, multi-infarct dementia and dementias of other aetiology according to ICD-10 and DSM-III-R. *Psychological Medicine* 1991;**21**:225–36.

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Brayne 1990

Study characteristics	
Patient Sampling	"The sample consisted of women aged between 70 and 79 selected from the age/sex register of a rural health centre in Cambridgeshire. This area was chosen for the stability of its population. Women only were studied to allow valid comparisons to be made between the age groups

Brayne 1990 *(Continued)*

	70-74 and 75-79. Two balanced strata were chosen from these age groups. The entire group aged 75-79 was approached (207), and of the 70-74 year olds 203 were randomly selected from 270 women. Ten subjects died before the interviewer (CB) was able to seem (sic) them, leaving 200 subjects available in each stratum"
Patient characteristics and setting	From Brayne and Calloway 1989a. "The final response rate of the available sample, excluding those who died before being interviewed, was 91%...44% and 45% of the younger and older groups respectively were in social class III, with 23% and 22% in social class IV; 17% of each group were in social class II, and just over 2% in social class I. Those in social class II and IV were mainly associated with agricultural occupations. Many women had also worked on the land themselves. Most of the women lived in independent accomodation, either with their spouse or on their own. Only two lived in long-stay hospitals, and two in residential homes. Fourteen per cent of the older group and 5% of the younger group lived in sheltered accomodation. More women in the younger group were married than widowed (53% and 42% respectively), whereas more women on the older group were widowed than married (58% and 36% respectively). Few were single or divorced. Educational level was similar for the two groups. In the younger group 79% and in the older group 73% left school at the statutory leaving age of 13/14 - although more of the older group (22%) than the younger group (4%) left school at 13. More of the older group (7%) than the younger group (3%) went on to some kind of further education"
Index tests	"The general practitioners were asked to rate whether they considered each subjected demented and/or depressed and if so to what degree. They were not asked whether the subject fulfilled any criteria, nor were they asked to consult their own notes. This was so the results would reflect their normal diagnostic practice"
Target condition and reference standard(s)	"Each subject was assessed for dementia, depression and other psychiatric disorders using CAMDEX diagnostic guidelines, which are similar to the Tenth International Classification of Diseases (WHO, 1998)"
Flow and timing	There is no evidence of partial verification, the interval between the index test and reference test is not specified. "Ten per cent of the interview schedules were randomly marked at the end, and at the termination of these interviews permission was requested for a further interview using GMS. Only one individual refused the GMS interview. The main interviewer was unaware until the end of the interview which subjects were to be asked. A consultant psychiatrist (PC), trained in the use of GMS, administered the hospital version of GMS within two weeks of the initial interview" Therefore all patients received CAMDEX which was the reference standard for dementia
Comparative	
Notes	Table 3 indicates that GPs rated 330 as "no dementia", from the text there are 311 TN: "in 311 both agreed that no dementia was present" so by subtraction there are 19 FN Table 3 indicates that GPs rated 29 as "dementia", from the text there are 10 TP: "the general practitioners' and clinical CAMDEX diagnoses agreed on dementia in 10 subjects" so by subtraction there are 19 FP

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		

Brayne 1990 *(Continued)*

Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it prespecified?	Unclear		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		

Brayne 1990 *(Continued)*

Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		Low risk

Cooper 1992

Study characteristics	
Patient Sampling	"In the German healthcare system, patients are not officially registered with individual GPs, but can change to a different doctor if they wish following each three-monthly period of care. Hence the sampling frame was based on consulting patients, rather than on a registered patient population. The doctor kept a record of all over-65-year-old patients that they saw during four weeks, whether at consultation or on home visits"
Patient characteristics and setting	"A total of 3737 patients were documented, but in 16 cases the information was seriously incomplete. The statistical analysis is based, therefore, on a collective of 3721 patients, with a mean age of 75.7 years (range 65-100 years). Of these 69.5% were women and 30.5% men, 91.8% were living in private households and 8.2% in long stay residential care"
Index tests	"In each case they (the doctors) made their own assessment of the patient's current level of cognitive functioning, with the help of simple guidelines supplied by the research team. The patient was allocated to one of four categories, as follows: *1 No impairment: normal memory, no difficulty in everyday activities, due to memory or cognitive deficits. *2 Mild forgetfulness: difficulty in recalling recent events and in unfamiliar tasks/ situations. Tendency to mislay and lose things. Still able to live independently. *3 Manifest impairment of attention and memory ("mild dementia"): Recent events and information forgotten at times. Occasional confusion or disorientation. Needs some help with everyday activities. *4 Severe memory loss and disorientation (moderate/severe dementia): Recent events wholly forgotten. Disorientated for time and place. Dependent on others for help with routine activities"
Target condition and reference standard(s)	"The global severity of dementia or any milder degree of cognitive disorder was rated on a five-point scale, according to the CAMDEX criteria"
Flow and timing	"In each practice we drew a stratified random sample for interview, containing equal numbers in each of the four categories (allocated by GP to no impairment, mild forgetfulness, mild dementia, moderate-severe dementia)... We then sought the agreement of the selected patients - or if they were severely demented that of their caregivers - to a home visit and interview by a doctor or psychologist from the research team, who had not been informed about the GPs ratings" A total of 3721 people were assessed by GPs of whom 507 were selected in the stratified sample and 407 were actually seen
Comparative	

Cooper 1992 *(Continued)*

Notes

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			

Cooper 1992 *(Continued)*

Was there an appropriate interval between index test and reference standard?	Unclear	
Did all patients receive the same reference standard?	No	
Were all patients included in the analysis?	No	
Could the patient flow have introduced bias?		Unclear risk

Creavin 2021

Study characteristics	
Patient Sampling	"Participants were people with symptoms of dementia, who were aged at least 70 years and had been referred by their GP to this research study. Symptoms of dementia were not specified but generally include disturbance in memory, language, executive function, behaviour, and visuospatial skills. Symptoms were required to be present for at least six months, and could be reported by the person themselves, a family member, a professional, or another person; there was no severity threshold. An accompanying informant was mandatory. All participants were offered free accessible transport and translation services. People were excluded if they had a known neurological disorder (i.e. Parkinsonism, Multiple Sclerosis, learning disability, Huntington's disease), registered blind, profound deafness (i.e. unable to use a telephone), psychiatric disorder requiring current secondary care input, or if cognitive symptoms were either rapidly progressive or co-incident with neurological disturbance. People with very severe dementia, operationalised as inability to consent, were excluded as they were judged by a lay advisory group to find the research process overly burdensome. GPs were encouraged to refer a consecutive series of all attending eligible patients to the study, regardless of their clinical judgement or any test results"
Patient characteristics and setting	"We recruited participants from 21 participating GP surgeries in the Bristol, North Somerset, and South Gloucestershire (BNSSG) area Research clinics were in four participating GP surgeries, strategically located for accessibility"
Index tests	"The referring GP recorded their clinical judgement using an electronic referral form during a consultation with their patient about cognitive symptoms. Clinical judgement was operationalised as normal, cognitive impairment not dementia (CIND), or dementia"
Target condition and reference standard(s)	Dementia ICD-10 criteria "At the research clinic, a single specialist physician with more than 20 years' experience in the field of dementia conducted a standardised assessment lasting approximately 60 minutes comprising clinical history, the Addenbrooke's Cognitive Examination III (ACE- III), Brief Assessment Schedule Depression Cards (BASDEC) and the Bristol Activities of Daily Living (BADL) Questionnaire. The specialist was not aware of other test results such as GP judgement or any investigations. The reference standard was based on the evaluation of the specialist physician for dementia according to ICD-10 criteria"
Flow and timing	All cases verified "The median time between referral (clinical judgement) and the clinic appointment (reference standard) was 47 days (IQR 30 to 72 days), the longest interval was 177 days, which was due to difficulties with attending earlier appointments"

Creavin 2021 *(Continued)*

Comparative			
Notes	Authors of this paper also authors on this review		
Methodological quality			
Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it prespecified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the refer-			Low concern

Creavin 2021 *(Continued)*

ence standard does not match the question?		
DOMAIN 4: Flow and Timing		
Was there an appropriate interval between index test and reference standard?	Yes	
Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		Low risk

De Lepeleire 2004

Study characteristics	
Patient Sampling	"During a one-month period between July and September 2000, during consultations in the office and during visits at home, they screened all subjects older than 65 years that met the inclusion criteria, with an expected minimum of 50 patients... By means of a case-finding strategy by general practitioners, subjects that possibly suffered from dementia were sought. All subjects older than 65 years that had contact with their general practitioner during the study period were screened. Exclusion criteria were: refusing to participate, living in a residential home for the elderly, a diagnosis of dementia and having a manifest handicap with an impact on IADL (as it could be a sign of complex morbidity especially cerebrovascular accidents). The determination of this criterion was based on the GP's clinical judgment"
Patient characteristics and setting	"The demographic data showed a preponderance of female patients (62.7%). The mean age was 75.1 years (SD 6.8, range 64–100). Half of the contacts were home visits, and the carer was included in one third (35.1%). One in three lived alone, but 60% lived with his/her partner"
Index tests	"Four-item IADL (Barberger-Gateau's four-Item IADL). The instrument gives a description of several levels of functioning for each of four items registered, numbered from one to five. After the IADL questions had been asked, the general practitioner had to give their opinion about whether the patient had dementia"
Target condition and reference standard(s)	"The specialist made the final judgment after evaluation of the results of the above mentioned instruments (Four-item IADL (Barberger-Gateau's four-Item IADL), Judgment of the general practitioner, Mini-Mental State Examination, Camdex-N) according to the DSM-IV criteria (American Psychiatric Association, 1994)"
Flow and timing	"In case of a score > 0 on the four item IADL, Folstein's Mini Mental State Examination was performed in a separate consultation, as a control group a MMSE was also performed in each tenth patient with a score of 0. In the case of a score of 4 on the IADL, the patient was referred for specialist examination including CT head-scan, Camdex-N, clinical and biochemical examinations. 1003 people were seen by the GP. 81.6% scored 0 on the IADL. The IADL scores of 1-4 were 8.6%, 4.8%, 2.8%, and 2.2% respectively. A MMSE was performed in 189 people: 124 with IADL > 0 and 65 from the control group. Eight people refused the MMSE. 10 underwent CAMDEX-N"

De Lepeleire 2004 *(Continued)*

Comparative			
Notes			
Methodological quality			
Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Unclear
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern

De Lepeleire 2004 *(Continued)*

DOMAIN 4: Flow and Timing		
Was there an appropriate interval between index test and reference standard?	Unclear	
Did all patients receive the same reference standard?	No	
Were all patients included in the analysis?	Unclear	
Could the patient flow have introduced bias?		High risk

Eefsting 1996

Study characteristics	
Patient Sampling	"The target population of the GP rating study consisted of all persons age 65 and over on 1 March 1991 and who were registered with eight GPs working in a rural area in the Netherlands" The GP study is consecutive, the community survey is random
Patient characteristics and setting	"At the start of the study, the total study population consisted of 2655 subjects over 65 years of age. Fifty-one former patients of the GPs were institutionalised in nursing homes or psychiatric hospitals, and therefore, were no longer the GPs' responsibility. Consequently, the population of the GP registration study consisted of 2604 subjects. The median age for the total study population was 73 years - 72 for men (range 65-98) and 73 for women (range 65-103)"
Index tests	"For all patients aged 65 and over, GPs had to indicate: (a) their acquaintance with the patient (patient known well enough to judge his cognitive state or not); (b) the number of consultations in the year prior to the assessment (0-3, 4-7 or > 7 visits); and (c) the cognitive status of the patient. For the latter judgement, the GPs were given a forced choice between dementia, cognitive impairment and no cognitive impairment. For the GP diagnosis of dementia, patients had to fulfil the diagnostic criteria of DSM-III-R for two reasons: (a) they were considered to be comprehensible; and (b) they were also used in the community survey. Unfortunately, no generally accepted criteria are available for the diagnosis of cognitive impairment. In the present study, the GP diagnosis of cognitive impairment refers to the situation where the GP has noted the presence of signs and symptoms indicative of the diagnosis dementia without being certain whether the patient meets all the necessary DSM-III-R criteria for that diagnosis. Rating took place simply on the basis of the GPs own recollections of their patients or available case records and directly preceded the invitation to the patients for the community survey"
Target condition and reference standard(s)	"Subjects entering the second stage were examined at home or in their place of residence by one of the first two authors, both are physicians with geriatric experience and trained to conduct the CAMDEX. The diagnosis was based on all information gathered during administration of the CAMDEX. For every subject diagnostic classification was applied according to DSM- III-R diagnostic criteria. Staging of dementia was done using DSM-III-R descriptions for mild, moderate and severe dementia as well as the CAMDEX-description of minimal dementia"
Flow and timing	GPs made a judgement about patients when they could. For the reference test the following procedure was followed: "In the first stage all patients were screened for the presence of cognitive impairment by a trained lay-interviewer over the course of 1 year using the Dutch version of the Mini- Mental State Examination... In order to obtain an enriched sample of cases, a non-proportional stratified random sample was drawn, based on the MMSE-score in the first stage. Subjects scoring 17

Eefsting 1996 *(Continued)*

	or below, together with a random two-out of-three sample of those scoring between 18 and 23 and a random one-out of-three sample of those scoring between 24 and 27 were invited for the second, diagnostic stage. No sample was taken of those scoring 28 and above, assuming that no cases of DSM-III-R dementia would be present in this stratum. A number of studies regarding the sensitivity of the MMSE for dementia (being 100% at the cutting point 27/28) support this assumption"
Comparative	
Notes	

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it prespecified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Unclear
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		

Eefsting 1996 *(Continued)*

Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		Unclear risk	

O'Connor 1988

Study characteristics	
Patient Sampling	"The names of all patients aged 75 and over on 1 April 1986 were taken from the age sex registers of six group practices in Cambridge city. A further one in three names were sampled from a seventh practice to make up the numbers required for a later part of the project. All the practices were accredited for training. Patients were asked through their doctors to take part in a screening interview which inquired into personal details, family contact, and health and concluded with the mini-mental state examination (MMSE). This brief, cognitive examination comprises tests of orientation, attention, language, and recall..." "Two groups of patients were excluded... those in long stay hospitals ..." and a "group of patients [who] were diagnosed as having minimal dementia"
Patient characteristics and setting	There is no information available on the age and sex and other demographics of the included participants
Index tests	"General practitioners and community nurses were asked to mark on a list of their patients which ones they considered to be definitely not demented, possibly demented, or definitely demented, but only those assessed with the CAMDEX are considered here"

O'Connor 1988 *(Continued)*

Target condition and reference standard(s)	"Dementia was diagnosed only when operational criteria appended to the examination protocol {CAMDEX} were satisfied"
Flow and timing	"Respondents who scored 23 or less on the mini mental state examination out of a maximum score of 30, together with a one in three sample of those who scored 24 or 25, were assessed in more detail by the CAMDEX"
Comparative	
Notes	

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	No		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Unclear		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern

O'Connor 1988 *(Continued)*

DOMAIN 4: Flow and Timing		
Was there an appropriate interval between index test and reference standard?	Unclear	
Did all patients receive the same reference standard?	No	
Were all patients included in the analysis?	No	
Could the patient flow have introduced bias?		High risk

Pentzek 2009

Study characteristics	
Patient Sampling	"The study cohort consists of all 3,242 patients participating in a prospective longitudinal study on early detection of MCI and dementia within the framework of the German Study on Ageing, Cognition and Dementia in Primary Care Patients (AgeCoDe). The patients were selected from the files of 138 GP surgeries in six metropolitan study centres (Hamburg, Bonn, Düsseldorf, Leipzig, Mannheim and Munich). Inclusion criteria were age between 75 and 89 years and at least one GP contact within the last 12 months. Main exclusion criteria were: diagnosis of dementia, non-regular patient of the GP surgery, GP contact by home visits only, residence in a nursing home, and severe illness probably leading to death within three months. After application of the inclusion and exclusion criteria, 40 to 60 patients were chosen at random from the list of the GP by the research staff and invited to participate in the study by a letter from their GP"
Patient characteristics and setting	"3,242 patients were included in the calculation of MCI prevalence. Their mean age was 80.2 (±3.6) years, 53.2% were in the age-group of 75-79 years, 37.3% in the group of 80-84 years and 9.5% in the group 85 years and older. 65.6% were female, 34.4% male. 62% hat a low level of education, 27.3 a middle level and 10.6% had a high level of education according to the CASMIN classification"
Index tests	"Those who reacted positively to the letter were invited to consult their GP and discuss the study with him. When a patient consented to participate hereafter, the GP documented the patients' diseases and his own familiarity with the patient. The GP also estimated the actual cognitive status of the patient on the Global Deterioration Scale (GDS) based on his general impression of the patient but without any structured testing"
Target condition and reference standard(s)	"The most important neuropsychological instrument used by the interviewers was the SIDAM (Structured Interview for the Diagnosis of Dementia of the Alzheimer Type, Multi-Infarct Dementia, and Dementias of other Aetiology according to DSM-IV and ICD-10)... MCI was identified according to new consensus criteria proposed by the International Working Group on Mild Cognitive Impairment (1). These criteria include: 1. absence of dementia according to DSM-IV or ICD-10, 2. evidence of cognitive decline: self-rating or informant report and impairment on objective cognitive tasks and/or evidence of decline over time on objective cognitive tasks, 3. preserved baseline activities of daily living or only minimal impairment in complex instrumental functions" The Winbland and Winblad-modified definitions of MCI are provided but data are only extracted for Winblad to allow comparison with other studies. The results of the Winblad modified definition were: TP 90 FP 136 FN 722 TN 2266 [Kaduszkiewicz] "The reference standard on dementia was constituted in consensus conferences with all interviewers and study coordinators. Suspected dementia cases and indefinite cases were reviewed using all available information (interview data, interviewer notes, informant information, and GP-documented diagnoses). A decision on disease status (dementia or not) and on dementia etiology was

Pentzek 2009 *(Continued)*

	derived according to Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, criteria" [Pentzek]
Flow and timing	It appears that all participants received both the index test and reference standard "GPs gave their judgments of their patients' cognitive status using the Global Deterioration Scale (GDS)2 at baseline and both follow-ups" Participants received the reference standard 1.5 and 3 years after the index test
Comparative	
Notes	

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			

Pentzek 2009 *(Continued)*

Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		High risk	

Pond 1994

Study characteristics	
Patient Sampling	"All GPs conducting clinics at a large retirement village complex in Sydney, Australia, were approached to take part in this study, provided that they had at least 10 patients aged 70 or over who were living either independently or in hostel accommodation. During the pre- and post-intervention sampling phases each GP completed a one page questionnaire for each patient attending and recorded his/her opinion on the dementia status of each of these patients, as well as an opinion on whether the patient was depressed or not. The reasons for attendance were recorded as were major chronic illnesses. Between 8 and 29 patients were seen by each GP over an average 4 week period" Data from the pre-intervention group are included
Patient characteristics and setting	Mean (SD) age 82.5 (5.9) years, 86% female, mean (SD) Mini Mental State Examination Score 26.0 (3.7)
Index tests	"Each GP completed a one page questionnaire for each patient attending and recorded his/her opinion on the dementia status of each of these patients"

Pond 1994 *(Continued)*

Target condition and reference standard(s)	"Patients who agreed to join the study were then, within a week of their consultation, interviewed in their own home by a registered nurse. A Mini-Mental State Examination was performed, the Geriatric Depression Scale administered and demographic information obtained. A one in two sub sample received an abridged version of the Canberra Interview for the Elderly (CIE), a structured interview with an informant component. The CIE enables the generation of a set of diagnoses for each patient based upon the DSM-III-R classification system and the draft International Classification of Diseases, version 10 (ICD-10). Selection for the CIE sub sample was random, taking every second consecutive case"
Flow and timing	"A one in two sub sample received an abridged version of the Canberra Interview for the Elderly (CIE), a structured interview with an informant component. The CIE enables the generation of a set of diagnoses for each patient based upon the DSM-III-R classification system and the draft International Classification of Diseases, version 10 (ICD-10). Selection for the CIE sub sample was random, taking every second consecutive case"
Comparative	
Notes	From table 2, disease positives = 33; n = 105 From table 4, 7 patients at baseline could not provide an accurate DSM-III-R diagnosis therefore the disease positives are 33-7 = 26 Assuming other metrics in table 4 are correct and that the disease positives = 26, then our calculated figure for TP + TN = 81 which is the same as the figure in table 4 for agreement between GP and CIE diagnosis. No other combination of figures works

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Unclear		

Pond 1994 *(Continued)*

Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Rondeau 2008

Study characteristics	
Patient Sampling	"GPs had to include in the trial five consecutive patients over 75 years old who suffered from a spontaneous memory complaint or who were reported to do so by an informant. The inclusion of the patients lasted from October 2002 to February 2004. For patients agreeing to participate in the study, signed informed consent was requested" Data are extracted for the control group only
Patient characteristics and setting	1005 women and 414 men; 89.6% resident at home, 9.3% resident in institution, 1.1% resident "other"; 25.8% no or primary education, 50.7% primary education, 23.5% secondary education or higher

Rondeau 2008 *(Continued)*

Index tests	"GPs from both the intervention and the control groups gave a diagnosis and classified all patients in one of three categories: suspected dementia, non-suspected dementia, uncertain"
Target condition and reference standard(s)	"The specialists used the NINCDS-ADRDA criteria and the Mini Mental State Examination to establish the diagnosis of AD"
Flow and timing	"Only subjects with suspected dementia who accepted referral to a specialist had confirmation of the diagnosis. From the non-suspected subjects not usually seen by a specialist, a sub-sample was seen because either the patient or his/her family wished for confirmation of the GP's diagnosis. For the uncertain group, the patient could be seen or not according to the GP's usual practice, or at the request of the patient or his/her family" Of 375 people diagnosed with dementia by GP, 222 were seen by a specialist Of 711 people not diagnosed with dementia by GP, 38 were seen by a specialist Of 311 people with an uncertain GP diagnosis, 125 were seen by a specialist
Comparative	
Notes	Data extracted from Table 2 in the original paper. TP (143) are number diagnosed with dementia by both GP and specialist, FP (79) are number diagnosed with dementia by GP and "no or possible dementia" by specialist, disease positives (202) are number diagnosed with dementia by specialist and disease negatives (197) are number diagnosed with "no or possible dementia" by specialist

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Unclear		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Unclear		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	

Rondeau 2008 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		High risk	

Valcour 2000

Study characteristics	
Patient Sampling	"The study took place in a 6-physician internal medicine out- patient practice within a predominantly Asian American community of Honolulu, Hawaii. Consecutive patients, aged 65 years or older and seen between August 17, 1998, and September 26, 1998, were invited to participate in a 1-hour interview of cognitive testing. Subjects identified a person (proxy) who could provide collateral history regarding memory and thinking. This was usually a family member (89%) who was often interviewed by telephone. All patients in the study lived in the United States; ethnicity was self-reported" "Six participants were not included in the final analysis due to the presence of factors thought to affect cognitive testing (depression [n = 2], severe aphasia [n = 1], severe hearing loss [n = 1], and the use of a medication that could affect cognition [n = 2])"
Patient characteristics and setting	From Table 1 of the paper: mean (SD) age: 74.6 (6.18) years; education mean (SD): 12.0 (3.07) years; female sex 63/297
Index tests	"Immediately after primary care physicians saw patients in their offices for a routine visit, they completed a form, which asked the following: “Based on this encounter and my previous experience with this patient, in my best opinion, does this patient have dementia?” Response options were “yes,” “no,” and “unsure.” Investigators were blind-

Valcour 2000 *(Continued)*

	ed to these responses until data analysis. Clinicians were aware that this question would be asked and had no study imposed restrictions regarding what tests they could perform during the office visits"
Target condition and reference standard(s)	"A single geriatrician (V.G.V.), involved in all testing, provided a diagnosis of dementia using Benson and Cummings' criteria, defined as acquired impairment in at least 3 of 5 domains, including memory, language, visuospatial ability, higher cognition, and mood or personality" The degree of impairment was rated by the Clinical Dementia Rating (CDR) scale
Flow and timing	All participants appear to have index test and reference test but no information is available on timing
Comparative	
Notes	Author of original paper contacted for information on FP and TN but data not available. Therefore test data not included in meta-analysis but included here. From Table 2 in the original paper, recognition of dementia at the time of the office visit: CDR mild (0.5 or 1) Recognised 1 Not recognised 10; CDR moderate (2) Recognised 2 Not recognised 2; CDR severe (> 2) Recognised 3 Not recognised 0; All CDR stages Recognised 6 Not recognised 12

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Unclear		

Valcour 2000 *(Continued)*

Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Wind 1995

Study characteristics	
Patient Sampling	"This study was part of the Amsterdam Study of the Elderly (AMSTEL), which investigates the course of cognitive functioning in elderly people in Amsterdam living at home. For our purposes we used cross-sectional data. For this study a two-stage sample was drawn from the practice lists of 30 general practices (21 selected randomly and nine by convenience) of 36 GPs in Amsterdam. First, a non-proportional age-stratified sample of patients was drawn; within each practice, subjects were randomly selected from each of four 5-year age strata (65-69 to 80-84) to arrive at equal-sized strata. Within each practice, 91-333 subjects were sampled and invited to participate in the screening study. The refusal rate was 27%; 4051 subjects consented to participate. In a non-response study we concluded that selection in refusal could not be excluded for the youngest age group (65-75 years). This possible selection on cognitive functioning is compensated by the second step: the score on the Mini-Mental State Examination (MMSE) from the screening study was used to select a sample of subjects for a diagnostic study. We selected all subjects who scored poorly on the MMSE (< 22), and randomly selected a sample of subjects who scored suspect or normal (between 22 and 26, and between 27 and 30). Individuals who were blind, deaf or non-Dutch-speak-

Wind 1995 *(Continued)*

	ing, as well as subjects with more than four missing values on the MMSE, were excluded. The refusal rate for the diagnostic study was 29%; 511 subjects finally participated. For 36 subjects, no GP judgement was available. Included in our study were 475 of the 51 1 community-dwelling subjects (65-84 years of age), who had a higher probability of having or developing dementia due to our selection"
Patient characteristics and setting	"There were 181 men (38%) in the study population and 294 women. The mean age was 75.2 years (SD = 5.7)"
Index tests	"The GP judgement on dementia was based on the GP's memory, using, if needed, the patient's record. Specially trained interviewers (three medical students) asked the general practitioners, during a practice visit, to categorize the cognitive functioning of every patient on an ordinal scale containing the categories of the gold standard. The diagnostic criteria of the gold standard were first explained to the GP. A score 'unknown' could be attributed if the GP did not remember any details about the patient.."
Target condition and reference standard(s)	"The gold standard in this study was the diagnosis of the Dutch version of the CAMDEX... Data on the dementia criteria were measured in a structured interview conducted by trained interviewers in the homes of the elderly participants. The interviewers were unaware of the GP judgements"
Flow and timing	All participants appear to have had both index and reference test but no information is available on timing
Comparative	
Notes	

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index Test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it prespecified?	Yes		

Wind 1995 *(Continued)*

Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Unclear
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

FN: false negative; FP: false positive; GPs: general practitioners; IQR: interquartile range; MCI: mild cognitive impairment; SD: standard deviation; TN: true negative; TP: true positive.

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Aldus 2018	Inappropriate index test: not gestalt clinical judgement
Belmin 2012	Inappropriate participants: not primary care
Borson 2006	Inappropriate participants: not primary care
Bushnell 2004	Inappropriate reference standard

Study	Reason for exclusion
Camicioli 2000	Inappropriate index test: not gestalt clinical judgement
Chong 2016	Inappropriate participants: not primary care
De Lepeleire 2005	Inappropriate index test: not gestalt clinical judgement
Dilts 2003	Inappropriate participants: not primary care
Dinesen 1997	Inappropriate index test: not gestalt clinical judgement
Engedal 1989	Inappropriate participants: not primary care
Fichter 1995	Inappropriate participants: not primary care
Hara 2013	Inappropriate reference standard
Hessler 2014	Inappropriate index test: not gestalt clinical judgement
Hopman-Rock 2001	Inappropriate reference standard
Jacinto 2009	Inappropriate participants: not primary care
Jansen 2007	Inappropriate reference standard
Juva 1994	Inappropriate index test: not gestalt clinical judgement
Kurz 1999	Inappropriate index test: not gestalt clinical judgement
Leung 2007	Inappropriate study design (i.e. not a diagnostic test accuracy study e.g. a study reporting qualitative data, descriptive epidemiology, randomised trial, or survey)
Lionis 2001	Inappropriate index test: not gestalt clinical judgement
Livingston 1990	Inappropriate index test: not general practitioner
Löppönen 2003	Inappropriate index test: not gestalt clinical judgement
Mant 1988	Inappropriate reference standard
Mok 2004	Inappropriate index test: not gestalt clinical judgement
Noda 2018	Inappropriate index test: not gestalt clinical judgement
Olafsdóttir 2000	Inappropriate index test: not gestalt clinical judgement
Pittmann 1992	Inappropriate participants: not primary care
Schaub 2003	Inappropriate index test: not general practitioner
Tierney 2014	Inappropriate reference standard
van Hout 1999	Inappropriate index test: not gestalt clinical judgement
van Hout 2000	Inappropriate index test: not gestalt clinical judgement

Study	Reason for exclusion
van Hout 2001	Inappropriate index test: not gestalt clinical judgement
van Hout 2002	Inappropriate index test: not gestalt clinical judgement
van Hout 2003	Inappropriate index test: not gestalt clinical judgement
van Hout 2006	Inappropriate index test: not gestalt clinical judgement
van Hout 2007a	Inappropriate index test: not gestalt clinical judgement
van Hout 2007b	Inappropriate index test: not gestalt clinical judgement
Waldorff 2005	Inappropriate reference standard
Wang 2017	Inappropriate index test: not gestalt clinical judgement
Wilkins 2007	Inappropriate participants: not primary care

DATA

Presented below are all the data for all of the tests entered into the review.

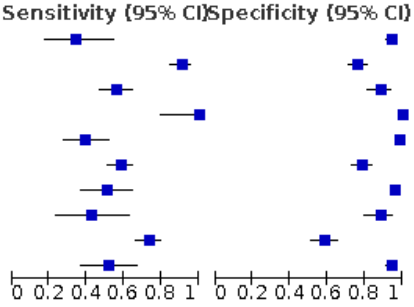
Table Tests. Data tables by test

Test	No. of studies	No. of participants
1 Clinical judgement for dementia	10	6087
2 Clinical judgement for cognitive impairment	5	4711

Test 1. Clinical judgement for dementia

Clinical judgement for dementia

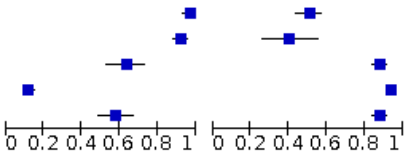
Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Brayne 1990	10	19	19	311	0.34 [0.18, 0.54]	0.94 [0.91, 0.96]
Cooper 1992	107	69	10	221	0.91 [0.85, 0.96]	0.76 [0.71, 0.81]
Creavin 2021	74	12	58	96	0.56 [0.47, 0.65]	0.89 [0.81, 0.94]
De Lepeleire 2004	16	0	0	987	1.00 [0.79, 1.00]	1.00 [1.00, 1.00]
Eefsting 1996	28	4	43	300	0.39 [0.28, 0.52]	0.99 [0.97, 1.00]
O'Connor 1988	121	51	87	185	0.58 [0.51, 0.65]	0.78 [0.73, 0.83]
Pentzek 2009	28	92	27	2147	0.51 [0.37, 0.65]	0.96 [0.95, 0.97]
Pond 1994	11	9	15	70	0.42 [0.23, 0.63]	0.89 [0.79, 0.95]
Rondeau 2008	143	79	52	111	0.73 [0.67, 0.79]	0.58 [0.51, 0.66]
Wind 1995	25	25	23	402	0.52 [0.37, 0.67]	0.94 [0.91, 0.96]



Test 2. Clinical judgement for cognitive impairment

Clinical judgement for cognitive impairment

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Cooper 1992	178	110	6	113	0.97 [0.93, 0.99]	0.51 [0.44, 0.57]
Creavin 2021	178	28	15	19	0.92 [0.88, 0.96]	0.40 [0.26, 0.56]
Eefsting 1996	62	33	36	244	0.63 [0.53, 0.73]	0.88 [0.84, 0.92]
Pentzek 2009	60	166	434	2554	0.12 [0.09, 0.15]	0.94 [0.93, 0.95]
Wind 1995	69	43	50	313	0.58 [0.49, 0.67]	0.88 [0.84, 0.91]



ADDITIONAL TABLES

Table 1. Circumstances for contacting authors to obtain information on diagnostic accuracy

Aspect of study that is not relevant to our review	Action we would take
Participants	Exclude the study
Reference standard	Exclude the study
Index test	Exclude the study
Target condition	Contact authors in the hope of obtaining information about diagnostic accuracy for target condition of interest only when we are confident from review of the full text that the participants, reference standard, and index test are applicable to the review. Where studies report the diagnostic accuracy of clinical judgement for the diagnosis of a composite target condition of cognitive impairment and dementia (e.g. cognitive impairment) we would attempt to obtain details of the diagnostic accuracy for each of our separate target conditions
Study design	Contact authors in the hope of obtaining information about diagnostic accuracy for target condition of interest only when we are confident from review of the full text that the participants, reference standard, index test, and target condition are applicable to the review

Table 2. Comparison of included citations in systematic reviews

Citation	Index test	Reference standard	Mitchell 2011	van den Dungen 2012	This review	Study ID in this review	Why excluded from this review
Boise 2004	Medical records	CERAD[a]	Yes	No	No	-	Index test
Borson 2006	Medical records	CERAD	Yes	No	No	-	Index test
Boustani 2005	Medical records	CERAD	Yes	No	No	-	Index test
Bowers 1990	Prospective questionnaire	MMSE[b]	Yes	No	No	-	Reference standard
Brayne 1990	Retrospective rating	CAMDEX[c]	No	No	Yes	Brayne 1990	-
Callahan 1995	Medical records	SPMSQ[d]	Yes	No	No	-	Reference standard; index test
Chodosh 2004	Medical records	TDQ[e]	Yes	No	No	-	Reference standard; index test
Cooper 1992	Prospective rating	CAMDEX	Yes	Yes	Yes	Cooper 1992	-
Creavin 2021	Prospective rating	ICD-10[f]	No	No	Yes	Creavin 2021	-
De Lepeleire 2004	Prospective rating	DSM-IV-TR[g]	No	No	Yes	De Lepeleire 2004	-
Eefsting 1996	Retrospective rating	CAMDEX	Yes	Yes	Yes	Eefsting 1996	-
Ganguli 2004	Medical records	MMSE	Yes	No	No	-	Reference standard; index test
Iliffe 1990	Medical records	MMSE	Yes	No	No	-	Reference standard; index test
Jacinto 2009	Medical records	Expert consensus	Yes	No	No	-	Reference standard; index test
Kaduszkiewicz 2010	Prospective rating	Mild cognitive impairment(Winblad 2004)	Yes	No	Yes	Pentzek 2009	-
Löppönen 2003	Medical records	DSM-IV-TR	Yes	Yes	No	-	Index test

Table 2. Comparison of included citations in systematic reviews *(Continued)*

Mant 1988	Doctors opinion	MMSE	Yes	No	No	-	Reference standard
O'Connor 1988	Retrospective rating	CAMDEX	Yes	No	Yes	O'Connor 1988	-
Olafsdóttir 2000	Medical records	DSM-III-R[h]	Yes	Yes	No	-	Index test
Pentzek 2009	Prospective rating	DSM-IV-TR	Yes	No	Yes	Pentzek 2009	-
Pond 1994	Prospective rating	DSM-III-R	Yes	Yes	Yes	Pond 1994	-
Rondeau 2008	Retrospective rating	DSM-IV-TR	Yes	No	Yes	Rondeau 2008	-
Valcour 2000	Prospective rating	CDR[i]	Yes	Yes	Yes	Valcour 2000	-
van Hout 2000	Prospective, applying Dutch guidelines	CAMDEX	Yes	No	No	-	Index test
Wilkins 2007	Medical records	CERAD	Yes	No	No	-	Index test
Wind 1995	Retrospective rating	CAMDEX	Yes	No	Yes	Wind 1995	-

[a]CERAD: Consortium to Establish a Registry for Alzheimer's Disease protocol (Morris 1993).
[b]MMSE: Mini-Mental State Examination (Folstein 1975).
[c]CAMDEX: Cambridge Mental Disorders of the Elderly Examination (Roth 1986).
[d]SPMSQ: Short Portable Mental Status Questionnaire (Pfeiffer 1975).
[e]Design aspect which did not meet criteria for current review. TDQ: Telephone Dementia Questionnaire (Kawas 1994).
[f]ICD-10: the International Classification of Diseases, 10th revision.
[g]DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders (4th edition, Text Revision).
[h]DSM-III-R: Diagnostic and Statistical Manual of Mental Disorders (3rd edition, Revised).
[i]CDR: Clinical Dementia Rating scale.

Table 3. Summary characteristics of included studies

Citation *Country*	Sampling	Index test	Reference standard	Flow and timing	Target condition	Definition	Access to medical records

Table 3. Summary characteristics of included studies *(Continued)*

Creavin 2021 *UK*	All people aged 70 years with cognitive symptoms attending 21 GP practices	Prospective	Expert	Verification complete	Dementia MCI	ICD-10 (ICD 1993)	Yes
O'Connor 1988 *UK*	All people aged 75 years and over from 6 GP lists in Cambridge and *1 in 3*[a] from a 7th surgery	Retrospective	CAMDEX	Timing not specified Partial verification: MMSE score < 24: all score 24 or 25: *1 in 3*[a]	Dementia	ICD-10	Unclear
Brayne 1990 *UK*	Randomly selected women aged 70 to 74 years, and all women aged 75 to 79 years from rural GP surgery list	Retrospective	CAMDEX	Timing not specified Verification complete	Dementia	ICD-10	No
Cooper 1992 *Germany*	Consulting patients over 65 years seen in 24 GP surgeries over 4 weeks	Prospective	CAMDEX	Timing not specified Partial verification: random sample stratified by GP opinion	MCI Dementia	ICD-10[b]	No
Wind 1995 *Netherlands*	Age stratified random sample from 30 GP surgery lists	Retrospective	CAMDEX	Timing not specified Partial verification: MMSE score < 22: all score > 22: random sample	Dementia	ICD-10	Yes
Pond 1994 *Australia*	Consulting patients in a retirement complex seen by GPs over 4 weeks	Prospective	Canberra Interview for the Elderly	Timing not specified Partial verification: 50% random sample	MCI Dementia	ICD-10[b]	Yes
Eefsting 1996 *Netherlands*	All patients aged 65 years and over on lists of 8 GPs	Retrospective	CAMDEX	Timing not specified Partial verification: MMSE score < 18: all score 18 to 23 random 2 in 3 score > 23 random 1 in 3 score > 27 none	MCI Dementia	DSM-III-R[b]	Yes

Table 3. Summary characteristics of included studies *(Continued)*

Valcour 2000 *USA*	Consecutive patients aged 65 years or more at 1 GP surgery over 6 weeks	Prospective	Expert	Timing not specified Flow unclear	Dementia	CDR	Yes
De Lepeleire 2004 *Belgium*	Consecutive patients aged 65 years or more with possible dementia were sought	Prospective	CAMDEX	Tests within 1 month Partial verification: IADL score 4: all seen score < 4 none seen	Dementia	DSM-IV-TR	Yes
Pentzek 2009 *Germany*	Random sample of people aged 75 to 89 years registered with GP and postal invitation to participate	Prospective	SIDAM	Reference test 1.5 years and 3 years after index test Complete verification	MCI Dementia	Winblad 2004 DSM-IV-TR	Yes
Rondeau 2008 *France*	Consecutive patients in a trial to train GPs. Only control patients included	Prospective	Expert	Timing not specified Partial verification: 222 of 375 diagnosed "dementia" by GP 38 of 711 diagnosed "not dementia" by GP 125 of 311 diagnosed "unsure" by GP	Dementia	NINCDS-ADRDA	Yes

CAMDEX: Cambridge Mental Disorders of the Elderly Examination; CDR: Clinical Dementia Rating scale; DSM-III-R: Diagnostic and Statistical Manual of Mental Disorders (3rd edition, Revised); DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders (4th edition, Text Revision); GP: general practitioner; IADL: Lawton instrumental activities of daily living scale; ICD-10: the International Classification of Diseases, 10th revision; MCI: mild cognitive impairment; MMSE: Mini-Mental State Examination; NINCDS-ADRDA: National Institute of Neurological and Communicative Diseases and Stroke/Alzheimer's Disease and Related Disorders Association criteria; SIDAM: structured interview for the diagnosis of dementia of the Alzheimer type, multi-infarct dementia and dementias of other aetiology according to ICD-10 and DSM-III-R.

[a]Further detail not reported.
[b]Same reference standard applies to both target conditions.

Brayne 1990 included the minimal cases on CAMDEX in with mild cases. Minimal dementia on CAMDEX approximates MCI.

Prospective index test: after consulting with a patient who has presented to a specific encounter with the doctor.

Retrospective index test: based on knowledge of the patient and review of the medical notes, but not relating to a specific encounter with the patient.

Table 4. Comparison of diagnostic accuracy studies

	Cooper 1992	Pond 1994	Valcour 2000	De Leeuw 2004	Pentzek 2009	Rondeau 2008	Brayne 1990	Eefsting 1996	Wind 1995	O'Connor 1988	Creavin 2021
Series[a]	C	C	C	C	R	C	R	C	R	C	C
Characteristics of participants											
Symptomatic[b]	No	No[c]	No	No	No	Yes	No	No	No	No	Yes
Mean age (years)	76	83	75	75	80	81	-[d]	73[e]	75	-	80
% female	70	86	63	63	66	71	100	-	62	-	47
% with dementia[e]	29	31	9	2	2	50	8	19	10	56	55
Verification with reference standard											
Verified (%)[e]	(11)	(53)	(100)	(1)	(70)	(26)	(100)	(15)	(100)	(100)	(100)
General practitioner (GP) judgement (%)											
Not impaired	36	-	33	-	94	48	90	90	76	45	14
Cognitive impairment no dementia (CIND)	41	-	-	-	-	-	7	8	62	20	40
Dementia	23	27	33	-	6	26	3	2	50	18	36
Uncertain	-	-	33	-	-	22	-	-	-	17	-
Diagnostic accuracy of GP judgement											
Sensitivity (%)	91	42	-	100	51	73	34	39	52	58	56
Specificity (%)	76	89	-	100	96	58	94	99	94	78	89

[a]C: consecutive, R: random, D: dementia.

Table 4. Comparison of diagnostic accuracy studies *(Continued)*

[b]Symptomatic: symptoms *required* for participation.

[c]Participants were not presenting with symptoms but GPs were asked to maximise the inclusion of people with suspected dementia.

[d]-: not reported.

[e]Median.

% verified =

number underwent reference test

number underwent index test

% with dementia =

number with dementia

number verified

P: prospective, B: retrospective.

APPENDICES

Appendix 1. Sources searched and search strategies

Source	Search strategy	Hits retrieved
MEDLINE In-process and other non-indexed citations and MEDLINE 1946 to present (Ovid SP) (Date of most recent search: 16 September 2021)	1. exp "sensitivity and specificity"/	Jan 2017: 5419
	2. "reproducibility of results"/	Jan 2018: 348
	3. diagnos*.ti.	Apr 2019: 927
	4. di.fs.	Nov 2020: 592
	5. sensitivit*.ab.	Sep 2021: 556
	6. specificit*.ab.	
	7. (ROC or "receiver operat*").ab.	
	8. Area under curve/	
	9. ("Area under curve" or AUC).ab.	
	10. sROC.ab.	
	11. accura*.ti,ab.	
	12. (likelihood adj3 (ratio* or function*)).ab.	
	13. ((true or false) adj3 (positive* or negative*)).ab.	
	14. ((positive* or negative* or false or true) adj3 rate*).ti,ab.	
	15. or/1-14	
	16. exp Dementia/	
	17. Neurocognitive Disorders/	
	18. dement*.mp.	
	19. alzheimer*.mp.	
	20. (lewy* adj2 bod*).mp.	
	21. (chronic adj2 cerebrovascular).mp.	
	22. ("organic brain disease" or "organic brain syndrome").mp.	
	23. ("normal pressure hydrocephalus" and "shunt*").mp.	
	24. "benign senescent forgetfulness".mp.	
	25. (cerebr* adj2 deteriorat*).mp.	
	26. (cerebral* adj2 insufficient*).mp.	
	27. (pick* adj2 disease).mp.	
	28. (creutzfeldt or jcd or cjd).mp.	
	29. huntington*.mp.	
	30. binswanger*.mp.	

(Continued)

31. korsako*.mp.

32. "cognit* impair*".mp.

33. exp *Cognition Disorders/

34. MCI.ti,ab.

35. ACMI.ti,ab.

36. ARCD.ti,ab.

37. SMC.ti,ab.

38. CIND.ti,ab.

39. BSF.ti,ab.

40. AAMI.ti,ab.

41. MD.ti,ab.

42. LCD.ti,ab.

43. QD.ti,ab.

44. AACD.ti,ab.

45. MNCD.ti,ab.

46. MCD.ti,ab.

47. ("N-MCI" or "A-MCI" or "M-MCI").ti,ab.

48. ((cognit* or memory or cerebr* or mental*) adj3 (declin* or impair* or los* or deteriorat* or degenerat* or complain* or disturb* or disorder*)).ti,ab.

49. "preclinical AD".mp.

50. "pre-clinical AD".mp.

51. ("preclinical alzheimer*" or "pre-clinical alzheimer*").mp.

52. (aMCI or MCIa).ti,ab.

53. ("CDR 0.5" or "clinical dementia rating scale 0.5").ti,ab.

54. ("GDS 3" or "stage 3 GDS").ti,ab.

55. ("global deterioration scale" and "stage 3").mp.

56. "mild neurocognit* disorder*".ti,ab.

57. (prodrom* adj2 dement*).ti,ab.

58. (episodic* adj2 memory).mp.

59. ("preclinical dementia" or "pre-clinical dementia").mp.

60. or/16-59

61. Family Practice/ or Ambulatory Care/

62. Physicians, Family/ or Physicians, Primary Care/

63. Primary Health Care/

64. "family practice*".ti,ab.

(Continued)

	65. "general practi*".ti,ab.	
	66. *General Practice/ or General Practitioners/	
	67. "family practitioner*".ti,ab.	
	68. "primary care".ti,ab.	
	69. Physician Assistants/	
	70. "physician assistant*".ti,ab.	
	71. Nurse Practitioners/ or Family Nurse Practitioners/	
	72. "nurse practitioner*".ti,ab.	
	73. or/61-72	
	74. 60 and 73	
	75. 15 and 74	
	76. "clinical judgement*".ti,ab.	
	77. "practitioner* judgement*".ti,ab.	
	78. ((clinician* or GP* or physician* or doctor*) adj3 (intuit* or recogni* or detect* or diagnos*)).ti,ab.	
	79. "gut feeling*".ti,ab.	
	80. gestalt.ti,ab.	
	81. "GP judgement*".ti,ab.	
	82. ((clinician* or GP* or physician* or doctor*) adj3 accura*).ti,ab.	
	83. *Practice Patterns, Physicians'/	
	84. or/76-83	
	85. 60 and 84	
	86. 75 or 85	
Embase 1974 to 15 September 2021 (Ovid SP) (Date of most recent search: 16 September 2021)	1. *diagnostic accuracy/ 2. reproducibility/ 3. diagnos*.ti. 4. sensitivit*.ab. 5. specificit*.ab. 6. (ROC or "receiver operat*").ab. 7. area under the curve/ 8. ("Area under curve" or AUC).ab. 9. sROC.ab. 10. accura*.ti,ab. 11. (likelihood adj3 (ratio* or function*)).ab.	Jan 2017: 2997 Jan 2018: 296 Apr 2019: 1010 Nov 2020: 607 Sep 2021: 3029

(Continued)

12. ((true or false) adj3 (positive* or negative*)).ab.
13. ((positive* or negative* or false or true) adj3 rate*).ti,ab.
14. "sensitivity and specificity"/
15. or/1-14
16. exp dementia/
17. Cognitive Defect/
18. dement*.mp.
19. alzheimer*.mp.
20. (lewy* adj2 bod*).mp.
21. (chronic adj2 cerebrovascular).mp.
22. ("organic brain disease" or "organic brain syndrome").mp.
23. ("normal pressure hydrocephalus" and "shunt*").mp.
24. "benign senescent forgetfulness".mp.
25. (cerebr* adj2 deteriorat*).mp.
26. (cerebral* adj2 insufficient*).mp.
27. (pick* adj2 disease).mp.
28. (creutzfeldt or jcd or cid).mp.
29. "cognit* impair*".mp.
30. huntington*.mp.
31. binswanger*.mp.
32. korsako*.mp.
33. MCI.ti,ab.
34. ACMI.ti,ab.
35. ARCD.ti,ab.
36. SMC.ti,ab.
37. CIND.ti,ab.
38. BSF.ti,ab.
39. AAMI.ti,ab.
40. LCD.ti,ab.
41. (QD or MD).ti,ab.
42. AACD.ti,ab.
43. MNCD.ti,ab.
44. MCD.ti,ab.
45. ("N-MCI" or "A-MCI" or "M-MCI").ti,ab.

(Continued)

46. ((cognit* or memory or cerebr* or mental*) adj3 (declin* or impair* or los* or deteriorat* or degenerat* or complain* or disturb* or disorder*)).ti,ab.

47. "preclinical AD".ti,ab.

48. "pre-clinical AD".ti,ab.

49. ("preclinical alzheimer*" or "pre-clinical alzheimer*").ti,ab.

50. (aMCI or MCIa).ti,ab.

51. ("CDR 0.5" or "clinical dementia rating scale 0.5").ti,ab.

52. ("GDS 3" or "stage 3 GDS").ti,ab.

53. ("global deterioration scale" and "stage 3").ti,ab.

54. "BSF".ti,ab.

55. "mild neurocognit* disorder*".ti,ab.

56. (prodrom* adj2 dement*).ti,ab.

57. (episodic* adj2 memory).ti,ab.

58. ("preclinical dementia" or "pre-clinical dementia").ti,ab.

59. or/16-58

60. general practice/ or General Practitioner/ or Family Nurse Practitioner/ or Clinical Practice/ or Primary Health Care/

61. ambulatory care/

62. primary medical care/

63. "family practice".ti,ab.

64. "general practi*".ti,ab.

65. "family practice*".ti,ab.

66. "family practitioner*".ti,ab.

67. "general practitioner*".ti,ab.

68. "primary care".ti,ab.

69. physician assistant/

70. "physician assistant*".ti,ab.

71. nurse practitioner/

72. "nurse practitioner*".ti,ab.

73. or/60-72

74. 15 and 59 and 73

75. "clinical judgement*".ti,ab.

76. "practitioner* judgement*".ti,ab.

77. ((clinician* or GP* or physician* or doctor*) adj3 (intuit* or recogni* or detect* or diagnos*)).ti,ab.

78. "gut feeling*".ti,ab.

(Continued)

	79. gestalt.ti,ab. 80. "GP judgement*".ti,ab. 81. ((clinician* or GP* or physician* or doctor*) adj3 accura*).ti,ab. 82. or/75-81 83. 59 and 82 84. 74 or 83	
PsycINFO 1806 to September week 2 2021 (Ovid SP) (Date of most recent search: 16 September 2021)	1. diagnos*.ti. 2. sensitivit*.ab. 3. specificit*.ab. 4. (ROC or "receiver operat*").ab. 5. area under the curve/ 6. sROC.ab. 7. accura*.ti,ab. 8. (likelihood adj3 (ratio* or function*)).ab. 9. ((true or false) adj3 (positive* or negative*)).ab. 10. ((positive* or negative* or false or true) adj3 rate*).ti,ab. 11. "sensitivity and specificity"/ 12. exp Test Reliability/ or exp Diagnosis/ or exp Medical Diagnosis/ 13. or/1-12 14. exp DEMENTIA/ or Neurocognitive disorders/ or Mild Cognitive Impairment/ or Cognitive Impairment/ 15. dement*.mp. 16. alzheimer*.mp. 17. (lewy* adj2 bod*).mp. 18. (chronic adj2 cerebrovascular).mp. 19. ("organic brain disease" or "organic brain syndrome").mp. 20. ("normal pressure hydrocephalus" and "shunt*").mp. 21. "benign senescent forgetfulness".mp. 22. (cerebr* adj2 deteriorat*).mp. 23. (cerebral* adj2 insufficient*).mp. 24. (pick* adj2 disease).mp. 25. (creutzfeldt or jcd or cid).mp. 26. huntington*.mp. 27. binswanger*.mp.	Jan 2017: 2043 Jan 2018: 152 Apr 2019: 367 Nov 2020: 221 Sep 2021: 385

(Continued)

28. korsako*.mp.

29. "cognit* impair*".mp.

30. MCI.ti,ab.

31. ACMI.ti,ab.

32. ARCD.ti,ab.

33. SMC.ti,ab.

34. CIND.ti,ab.

35. BSF.ti,ab.

36. AAMI.ti,ab.

37. MD.ti,ab.

38. LCD.ti,ab.

39. QD.ti,ab.

40. AACD.ti,ab.

41. MNCD.ti,ab.

42. MCD.ti,ab.

43. ("N-MCI" or "A-MCI" or "M-MCI").ti,ab.

44. ((cognit* or memory or cerebr* or mental*) adj3 (declin* or impair* or los* or deteriorat* or degenerat* or complain* or disturb* or disorder*)).ti,ab.

45. "preclinical AD".ti,ab.

46. "pre-clinical AD".ti,ab.

47. ("preclinical alzheimer*" or "pre-clinical alzheimer*").ti,ab.

48. (aMCI or MCIa).ti,ab.

49. ("CDR 0.5" or "clinical dementia rating scale 0.5").ti,ab.

50. ("GDS 3" or "stage 3 GDS").ti,ab.

51. ("global deterioration scale" and "stage 3").ti,ab.

52. "BSF".ti,ab.

53. "mild neurocognit* disorder*".ti,ab.

54. (prodrom* adj2 dement*).ti,ab.

55. (episodic* adj2 memory).ti,ab.

56. ("preclinical dementia" or "pre-clinical dementia").ti,ab.

57. or/14-56

58. exp Family Physicians/ or exp Primary Health Care/

59. exp General Practitioners/ or exp Clinical Practice/

60. "family practice".ti,ab.

61. "general practi*".ti,ab.

(Continued)

	62. "family practices".ti,ab. 63. "family practitioner*".ti,ab. 64. "general practitioner*".ti,ab. 65. "primary care".ti,ab. 66. "physician assistant*".ti,ab. 67. "nurse practitioner*".ti,ab. 68. or/58-67 69. 13 and 57 and 68 70. "clinical judgement*".ti,ab. 71. "practitioner* judgement*".ti,ab. 72. ((clinician* or GP* or physician* or doctor*) adj3 (intuit* or recogni* or detect* or diagnos*)).ti,ab. 73. "gut feeling*".ti,ab. 74. gestalt.ti,ab. 75. "GP judgement*".ti,ab. 76. ((clinician* or GP* or physician* or doctor*) adj3 accura*).ti,ab. 77. or/70-76 78. 57 and 77 79. 69 or 78	
Web of Science core collection (Date of most recent search: 16 September 2021)	TOPIC: (dement* OR alzheimer* OR "lewy bod*" OR DLB OR "vascular cognitive impairment*" OR FTD OF FTLD OR "cerebrovascular insufficienc*" OR "mild cognitive impairment" OR MCI) AND TOPIC: ("primary care" OR "general practi*" OR GP OR "doctor* surgery" OR "family practi*" OR "ambulatory care") AND TOPIC: (diagnosis OR sensitiv* OR specificit* OR ROC OR "receiver operat*" OR "Area under curve" or AUC OR sROC OR accura* OR "follow*-up" OR "positive predictive value*" OR "negative predictive value*" OR longitudinal OR longitudinally) AND TOPIC: (((GP OR practitioner* OR clinician*) AND (judgement* OR assessment* OR diagnosis)) OR "gut feeling*" OR gestalt) Timespan: All years. Search language=Auto	Jan 2017: 585 Jan 2018: 53 Apr 2019: 210 Nov 2020: 130 Sep 2021: 349
LILACS (BIREME) (Date of most recent search: 16 September 2021)	dementia OR demencia OR demência OR alzheimer OR alzheimers OR alzheimer's OR cognition OR "mild cognitive impairment" [Words] and "primary care" OR "general practice" OR "atención primaria" OR "Práctica general" OR "Prática geral" [Words]	Jan 2017: 38 Jan 2018: 0 Apr 2019: 13 Nov 2020: 8 Sep 2021: 11
TOTAL before de-duplication		Jan 2017: 11,082 Jan 2018: 849 Apr 2019: 2527

(Continued)

	Nov 2020: 1550 Sep 2021: 4330 **TOTAL: 18,201**
TOTAL after de-duplication	Jan 2017: 8045 Jan 2018: 501 Apr 2019: 1058 Nov 2020: 923 Sep 2021: 3643 **TOTAL: 12,427**

Appendix 2. Anchoring statements for assessment of risk of bias using QUADAS-2

Selection	Index test	Reference standard	Flow
Was a consecutive or random sample of patients enrolled? [yes/no] Consecutive or random sampling from patients in primary care would be considered at low risk of bias	*Were the index test results interpreted without knowledge of the results of the reference standard? [yes/no]* Studies at low risk of bias are likely to use terms such as 'blinded' or 'masked.' Studies that do not explicitly state that access to medical records was denied would be judged as unclear. See Index tests	*Is the reference standard likely to correctly classify the target condition? [yes/no]* See Reference standards. We would only include studies that use a recognised research definition of dementia or cognitive impairment which we would judge at low risk of bias	*Was there an appropriate interval between index test(s) and reference standard? [yes/no]* A study with an average delay between assessments of 6 months or less would be judged at low risk of bias. A study with a average delay of more than a year would be judged at high risk of bias. For delayed follow-up as a reference standard, follow-up should occur at least 3 months after the index test assessment
Was a case-control design avoided? [yes/no] We would not include case-control studies	*If a threshold was used, was it prespecified? [yes/no]* See Data extraction and management. There is no accepted cut-point for the index test. This item is likely to be of limited value in this review	*Were the reference standard results interpreted without knowledge of the results of the index test? [yes/no]* Studies at low risk of bias are likely to use terms such as 'blinded' or 'masked.' Studies that state that the reference standard assessment was allowed knowledge of the index test would be judged as high risk. Many studies may be at unclear risk of bias in this domain because of the possibility of referral letters from general practitioners to specialists	*Did all patients receive a reference standard? [yes/no]* Many studies in primary care that are not primarily designed as prospective research studies may be at high risk of bias in this domain. See Index tests

(Continued)

		Cross-sectional studies may be at higher risk of bias in this domain unless masking is explicit	
Did the study avoid inappropriate exclusions? [yes/no] Example of high risk of bias would be exclusions based solely on age, educational attainment, or place of residence. Example of low risk of bias would be terminally ill people	-	-	*Did all patients receive the same reference standard? [yes/no]* It is likely that at least some participants would not receive the reference standard in all studies
-	-	-	*Were all patients included in the analysis?* *[yes/no]* A maximum proportion of dropouts to remain low risk of bias has been specified as 20%
Could the selection of patients have introduced bias? [High/low/unclear] If exclusions were not explicit in the article or after contacting authors, we would judge this as unclear. Studies at high risk of bias would often use a sampling method that is not consecutive or random and/or exclude people inappropriately	*Could the conduct or interpretation of the index test have introduced bias?* *[High/low/unclear]* See Index tests. We propose that the core feature of clinical judgement is that it is unaided by any additional test, investigation or inquiry beyond that which is immediately available to the clinician. Provided that the index test meets the definition we use, the risk of bias for this item may be low risk. However, if it is not explicit that no other brief cognitive tests were used, then the item may be at unclear risk of bias	*Could the reference standard, its conduct, or its interpretation have introduced bias?* *[High/low/unclear]* Even allowing for an acceptable reference standard, studies may often be at unclear risk of bias in this domain unless it is explicit that the reference standard was applied independently of the index test	*Could the patient flow have introduced bias? High/low/unclear]* Many studies that are not primarily designed as research studies are likely to be at high risk of bias in this domain
Are there concerns that the included patients do not match the review question? Studies with high applicability would commonly include frail elderly people with multi-morbidity. Studies with low applicability would exclude these people. Studies with a prevalence of dementia or cognitive impairment of more than 70% would often be of low applicability	*Are there concerns that the index test, its conduct, or interpretation differ from the review question?* See Index tests. So long as the clinical judgement about dementia or cognitive impairment had been made by a primary care physician/general practitioner, we would judge this at high applicability	*Are there concerns that the target condition as defined by the reference standard does not match the review question?* So long as the reference standard was one of our listed definitions, we would judge this at high applicability	-

HISTORY

Protocol first published: Issue 2, 2017

CONTRIBUTIONS OF AUTHORS

All authors contributed to the manuscript and approved the submitted version.

DECLARATIONS OF INTEREST

Four review authors (ST Creavin, Y Ben-Shlomo, S Purdy, and S Cullum) are authors on the included Creavin 2021 paper. A neutral assessment of the study was guaranteed by data extraction and quality assessment for this study being done by two different review authors who are not authors on that paper.

SOURCES OF SUPPORT

Internal sources

- No sources of support provided

External sources

- NIHR, UK

  This review was supported by the National Institute for Health and Care Research (NIHR), via Cochrane infrastructure funding to the Cochrane Dementia and Cognitive Improvement group. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the Evidence Synthesis Programme, NIHR, National Health Service or the Department of Health and Social Care.

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

There were five differences between the methods as described and those in the published protocol. Firstly, studies were not restricted to people with cognitive symptoms because a) this restriction was judged on reflection to be incompatible with the retrospective definition of clinical judgement (which was included in the published protocol), and b) this approach would have been overly restrictive. Secondly, all abstracts were screened by two review authors rather than one. Thirdly, information on covariates was not extracted for stage of dementia, experience of general practitioners, proportion of male and female doctors, or type of practice because these factors were found to be poorly reported and judged on reflection to be of less relevance to the review question. Fourthly, based on discussion with expert statistical advisers, the main analysis was in studies at lowest risk of bias and a sensitivity analysis was done when studies at high risk of bias in two or more QUADAS-2 domains were included. Finally, we did not search ALOIS as this resource was no longer suitable for searches for diagnostic test accuracy studies, and we did not search BIOSIS as it was felt that the Web of Science Core Collection was an adequate source from the Web of Science platform for this topic area.

INDEX TERMS

Medical Subject Headings (MeSH)

*Alzheimer Disease [diagnosis]; Clinical Reasoning; *Cognitive Dysfunction [diagnosis]; Cross-Sectional Studies; *Dementia [diagnosis]; *Physicians, Primary Care; Prospective Studies; Retrospective Studies; Sensitivity and Specificity

MeSH check words

Aged; Female; Humans