Interrater agreement in dementia diagnosis: A systematic review and meta-analysis

Enzo Cerullo1,2 | Terry J. Quinn3,2 | Jenny McCleery4 | Elpida Vounzoulaki5 | Nicola J. Cooper1,2 | Alex J. Sutton1,2

1Department of Health Sciences, Biostatistics Research Group, University of Leicester, Leicester, UK
2NIHR Complex Reviews Support Unit, University of Leicester & University of Glasgow, Glasgow, UK
3Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK
4Oxford Health NHS Foundation Trust, Elms Centre, Banbury, UK
5Diabetes Research Centre, University of Leicester, Leicester General Hospital, Leicester, UK

Correspondence
Enzo Cerullo, Department of Health Sciences, Biostatistics Research Group, University of Leicester, Leicester LE1 7RH, UK.
Email: enzo.cerullo@bath.edu

Funding information
National Institute for Health Research, Grant/Award Number: 10.13039/501100000272

Abstract

Objectives: Dementia remains a clinical diagnosis with a degree of subjective assessment and potential for interrater disagreement. We described interrater agreement of clinical dementia diagnosis for various diagnostic criteria.

Methods: We conducted a PROSPERO-registered (CRD42020168245) systematic review and meta-analysis. We searched multiple cross-disciplinary databases from inception until April 2020 for relevant papers, extracted data and described study quality in duplicate. Study quality was assessed using the Guidelines for Reporting Reliability and Agreement Studies. We used random-effects models to obtain summary estimates of interrater agreement using kappa and, where possible, Gwet's AC1/2 coefficients.

Results: We found 7577 titles and 22 eligible studies. Meta-analysis was possible for all-cause dementia using the Diagnostic and Statistical Manual of Mental Disorders third edition revised (DSM-III-R) criteria (kappa = 0.66, 95% CI = [0.53,0.78]), Alzheimer's disease using the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association (NINCDS-ADRDA) criteria (kappa = 0.71, 95% CI = [0.65,0.77] for presence/absence and AC2 = 0.61, 95% CI = [0.53,0.70] when distinguishing probable/possible cases), and vascular dementia using the International Classification of Diseases version 10 (ICD-10) criteria kappa = 0.79 (95% CI = [0.70,0.87]). Data was more limited for other criteria and dementia types. AC1/2 coefficients generally indicated higher agreement. One study was rated as high quality.

Conclusions: Diagnostic criteria for clinical dementia may have good but imperfect agreement. This has important implications for clinical practice and research studies, which frequently assume these criteria are perfect tests, such as diagnostic test accuracy studies frequently conducted for biomarkers and neuropsychological tests, and for trials where incident dementia is the outcome.

KEYWORDS
dementia, diagnosis, interrater agreement, meta-analysis, reliability, systematic review

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. International Journal of Geriatric Psychiatry published by John Wiley & Sons Ltd.
1 | INTRODUCTION

Dementia may be diagnosed at two levels: either as the syndrome of dementia (irrespective of cause) or at the level of subtypes defined by the underlying neuropathology. The diagnostic criteria used clinically to identify dementia subtypes are largely based on symptomatic variations, and are the end result of progressive (and ongoing) refinement of clinico-pathological correlations. These correlations remain imperfect, especially in the oldest old, in whom post-mortem neuropathological examination cannot always distinguish those with and without cognitive impairment in life. Revisions of diagnostic criteria aim to improve interrater agreement and to incorporate new knowledge gained through research, particularly through biomarkers with the aim of eventually being able to identify the pathologies underlying individual cases of dementia.

A variety of validated diagnostic criteria are available. Many of these have been superseded by newer diagnostic criteria, but others—established by different organisations—are in concurrent use. There is typically some uncertainty in the diagnosis of dementia and its subtypes, which is reflected in the fact that many diagnostic criteria for dementia subtypes allow ‘probable’ and ‘possible’ diagnoses depending on how closely the patient’s symptoms conform to the archetypal case. These diagnostic criteria are often used and referred to as reference standard (or gold standard) tests. In particular, they often serve as reference standards in studies of the diagnostic test accuracy (DTA) of neuropsychological assessments and of biomarkers.

Although diagnostic criteria seek to operationalize dementia assessment, there remains an element of subjectivity. Thus, a potential source of imperfection in these diagnostic criteria is suboptimal interrater agreement. This is defined as the degree to which two or more raters make the same diagnosis under similar assessment conditions. Imperfect interrater agreement is an important potential source of error in the diagnostic criteria, and could cause several issues. For example, it could lead to inaccurate diagnosis which will then lead to inappropriate treatment in clinical practice. It could also cause the estimates from DTA studies to be over- or under-estimated. DTA studies are carried out to determine the accuracy and appropriate threshold of cognitive assessments and biomarkers, so the implications of biased estimates could be significant. There could also be bias for randomised controlled trials (RCTs). For instance, prevention studies will use dementia incidence as an outcome; imperfect interrater agreement will lead to misclassification, which means that the study could require a larger sample size to show an effect. Another issue arising from imperfect agreement is bias in the estimates of inter-criteria studies. These are studies which assess the agreement between different diagnostic criteria, to determine whether they identify the same groups of patients. Imperfect interrater agreement could cause the results of these studies to be questionable. Finally, imperfect interrater agreement could also lead to bias in the estimates of prevalence rates of clinical dementia.

A systematic review and meta-analysis has not been previously conducted on this topic. The aim of this systematic review and meta-analysis is to determine the interrater agreement for dementia diagnostic criteria.

2 | METHODS

Where appropriate, we followed the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) best practice guidelines for design, conduct and reporting of the systematic review. We worked according to a predefined protocol (PROSPEROPRINT) CRD42020168245) which is registered at https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=168245.

2.1 | Search strategy and inclusion criteria

We searched MEDLINE (Ovid), EMBASE (Ovid), PsycINFO (EBSCO) and CINAHL (EBSCO) for relevant studies until April 2020. We created a search syntax based on concepts of dementia and interrater agreement and used a validated search filter for dementia. The full search strategy is provided in Table S1.
All aspects of searching, data extraction and analyses were performed by two raters (EC and EV) working independently. At each stage results were compared, and consensus reached, with recourse to a third reviewer if required. Titles and abstracts generated from the search were screened to determine relevancy. Full-text articles were screened to determine eligibility. The inclusion/exclusion criteria were: (1) Studies having measured and quantified interrater agreement for at least one clinical dementia criteria (note that for this paper, this includes studies which have a short (≤2 weeks) time interval between measurements, sometimes called test-retest reliability studies, as long as they used different raters), (2) the dementia subtype(s) studied were clearly stated, (3) the diagnostic criteria used were clearly stated, (4) the study was published in English Language. Conference proceedings, clinical guidelines, dissertations, as well as letters and commentaries were excluded. Reference lists of included studies were hand-searched for additional eligible studies. Study authors were contacted to obtain the full-text when it was not available.

2.2 | Diagnostic criteria for dementia

For all-cause dementia, the diagnostic criteria considered for inclusion were the iterations of the International Classification of Diseases (e.g., ICD-10), the iterations of the DSM (e.g., DSM-III-R, DSM-IV, DSM-5). For Alzheimer’s dementia (AD), diagnostic criteria included the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s disease and Related Disorders Association (NINCDS-ADRDA), the NIA-AA core clinical criteria, and the various iterations of the DSM and ICD diagnostic criteria. For VaD, they included the National Institute of Neurological Disorders and Stroke and Association Internationale pour la Recherche et l’Enseignement en Neurosciences (NINDS-AIREN), and the various iterations of the ICD and DSM criteria. For Frontotemporal dementia (FTD), they included the Lund-Manchester criteria, the various DSM and ICD iterations, and the international behavioural variant TD criteria consortium (FTDC) criteria. Finally, for Dementia with Lewy bodies (DLB) we included various iterations of the McKeith criteria and DLB consortiums.

2.3 | Study quality assessment

Study quality was assessed using criteria based on the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). There were 10 items, where each study could be awarded a maximum of one point for each of the items. If a study scored 4 or less, study quality was deemed as low, between 5 and 7 moderate, and 8–10 high. The 10 items assessed were dementia reference standard criteria, dementia subtype, assessor population, sample size calculation, sampling methods, blinding between raters (rater-rater binding), whether raters were aware of patients’ previous diagnoses (rater-patient binding), assessment timing, and whether agreement was described with a measure of uncertainty. If enough data were available, sensitivity analyses limited to studies classified as high quality was planned.

2.4 | Data extraction and statistical analyses

Data extraction was independently undertaken by two authors (EC and EV) according to The Cochrane Handbook guidelines. Findings were reported according to PRISMA guidance, and any disagreements were settled by consensus between authors. For study characteristics, we extracted data for the number of subjects used for the assessment of interrater agreement, number and descriptions of raters, country the study was conducted in, whether the study was conducted in a single site, dementia type(s) evaluated, diagnostic criteria evaluated, estimated dementia prevalences, study settings, age information, gender, education information, ethnicity, whether any of the raters’ diagnoses were made face-to-face with the patient, information on the severity of dementia, sampling information.

Ethnicity was defined as ‘white,’ ‘non-white’ or ‘mixed,’ where studies reporting a population consisting of at least 80% of the same ethnicity were classified as ‘white’ or ‘non-white.’ Those studies that reported a population consisting of less than 80% ‘white’ or ‘non-white’ were classified as ‘mixed.’ If ethnicity was not reported for a study, it was defined based on the predominant ethnicity of the country(where the study was conducted).

Where reported, information on severity of cognitive decline was also recorded. Studies were labelled as ‘harder-to-classify’ if they reported over two-thirds of patients belonged to categories which indicate they were more difficult to classify (i.e., possible, mild dementia or MCI). Conversely, studies were labelled as ‘easier-to-classify’ if over two-thirds of patients were in categories which indicate that they were more straightforward to classify, such as probable dementias, severe dementia and healthy controls. Dementia severity was defined as ‘mixed’ if less than two-thirds of patients fitted into either of the aforementioned categories. We used reported study information for this where available. If this was not available, we used the mean classifications from the raters in the study.

Agreement was primarily described using a measure of agreement called the kappa-statistic, which is used to measure inter-rater agreement and takes into account the possibility of the agreement occurring by chance. For comparisons which were ordinal (e.g., probable vs. possible vs. no dementia) using 2 raters, we used weighted kappa-statistics with linear weights to account for the ordered structure, if possible. In addition, for studies which supplied sufficient data, we calculated Gwet’s AC1 and linearly-weighted AC1 (i.e., Gwet’s AC2) statistics using the irrCAC package in R and compared them to the kappa-statistics. For kappa-statistics and AC1/2, we make reference to the classifications according to Altman, corresponding to ‘poor’ (0.20 or less), ‘fair’ (0.21–0.40), ‘moderate’ (0.41–0.60), ‘good’ (0.61–0.80), and ‘very good’ (0.81 and over) agreement.

Estimates were tabulated by diagnostic criteria, study and comparisons being made. Where at least four estimates were available,
kappa-statistics were pooled by dementia type, diagnostic criteria and comparison, by fitting random-effects meta-analytic models in R using metafor packages, using the method described by Sun et al. For these pooled estimates we presented both confidence intervals (CIs) and prediction intervals (PIs), the latter better reflects the variation across different settings in the presence of heterogeneity, and what the expected estimate is for a future study. If possible, we also repeated these meta-analyses using AC1/2 coefficients using the same method. Any estimates which were not pooled were summarised in a narrative. If studies reported more than one assessment (e.g., before and after a standardisation meeting), then we used the first assessment. If studies did not report confidence intervals (CIs), where possible, they were calculated in R with packages fmsb and boot using the standard error, observed agreement or contingency tables. The proportion of variability that was due to between-study heterogeneity rather than within-study variability was assessed using the I²-squared statistic. A 95% CI was given, rather than only a point estimate, since the I²-squared statistic is known to be biased unless a large number of studies is available.

For the main meta-analysis, we pooled kappa-statistics for the presence or absence of dementia, by dementia type and diagnostic criteria. We also conducted a second meta-analysis looking at agreement for comparisons which take the uncertainty of diagnosis into account. To determine agreement for a given dementia type, we also conducted exploratory meta-analysis for each dementia type regardless of classification system used. If a sufficient number of studies were available, we planned to perform subgroup analyses for any pooled estimates obtained by dementia severity category, ethnicity and study settings, as well as meta-regression models for any pooled estimates obtained to investigate sources of between-study heterogeneity. If possible, publication bias was assessed using funnel plots for asymmetry using the Beggs and Eggers tests. Sensitivity analysis was also planned, to explore the effect of each individual study on overall pooled estimates.

3 | RESULTS

3.1 | Identified studies

A total of 7577 titles were screened (Figure 1). Sixty-nine full texts were assessed for eligibility. Twenty-five of these studies did not assess interrater agreement and 13 did, however this was not evaluated for clinical dementia. Six studies assessed interrater agreement for clinical dementia, but not for one of the specified diagnostic criteria. For two studies, the diagnostic criteria used was not clear, and for one study (Kukull 1990) we could not obtain the full text after contacting the authors. Hogervorst 2003 and Hogervorst 2000 both used patients from the Oxford project to investigate Memory and Ageing (OPTIMA) cohort. To avoid multiple uses of the same patients, we used data from the latter study to obtain a pooled estimate for the NINCDS-ADRDA due to a larger sample size. A total of 22 studies met the eligibility criteria.

3.2 | Study characteristics

The studies included in the systematic review investigated different dementia subtypes and used various diagnostic criteria for dementia (Table 1 and Table 2). They had a wide range of estimated dementia prevalences (0.03–1.00). Thirteen studies reported sufficient information to estimate dementia severity. Of these, 5 were easier to classify, 7 were of mixed, and one more difficult to classify. Eight studies reported information on education. One reported that 84% of participants had education level ‘greater than high school,’ 1 ‘at least high school,’ 6 reported mean or median education levels of at least 10 years, and 1 reported a mean education of only 4.9 years. Nineteen studies were conducted in white ethnicity populations and three mixed. Thirteen studies were conducted in secondary care settings, five in community settings and four using both secondary and community settings. Fourteen studies reported sampling information; three used random sampling, three stratified, six stratified, six convenience, and two consecutive. Nine studies were conducted using a single site, and 16 studies had more than two raters.

3.3 | Study quality

All studies which underwent quality assessment using the GRRAS obtained a rating between 2 and 9. 11 studies were rated as a low study quality, 10 moderate and 1 high. Assessment of quality is described in Table S3.

3.4 | Meta-analysis

3.4.1 | Analysis by diagnostic criteria & dementia type (main analysis)

The first meta-analysis investigated the interrater agreement of the DSM-III-R for the presence or absence of all-cause dementia. It included five studies and we obtained a pooled kappa-statistic of 0.66 (95% CI = [0.53, 0.78], 95% PI = [0.40, 0.92]; Figure 2). The second meta-analysis included eight studies assessing and investigated the interrater agreement of the NINCDS-ADRDA for AD, where probable and possible AD groups were merged. The kappa-statistic was found to be 0.71 (95% CI = [0.65, 0.77], 95% PI = [0.65, 0.77]; Figure 2). The third meta-analysis investigated the interrater agreement of the ICD-10 for presence or absence of VaD and included four studies. We obtained a kappa-statistic of 0.79 (95% CI = [0.70, 0.87], 95% PI = [0.70, 0.87]; Figure 2).

We conducted another analysis investigating the agreement which takes the uncertainty of diagnosis into account. We were able to do this for the NINCDS-ADRDA criteria for AD for Pr-AD versus Ps-AD versus non-AD (Figure 3). We obtained a weighted kappa-statistic of 0.59 (95% CI = [0.51, 0.67], 95% PI = [0.48, 0.70]). For the presence or absence of AD for these 4 studies, we obtained a kappa-statistic of 0.70 (95% CI = [0.61, 0.79], 95% PI = [0.60, 0.90]).
and an I-squared statistic of 6 (95% CI = [0, 95]). We were also able to conduct this analysis using AC1/2 statistics (Figure 4). Compared to the analysis using kappa-statistics, we obtained a very similar agreement for Pr-AD versus Ps-AD versus non-AD AC2 = 0.61 (95% CI = [0.53, 0.70], 95% PI = [0.49, 0.74]), and for the presence or absence of AD we obtained a higher agreement AC1 = 0.75 (95% CI = [0.65, 0.86], 95% PI = [0.59, 0.92]). For other diagnostic criteria and dementia subtypes, there was not enough data to obtain pooled estimates.

3.4.2 | Exploratory analyses

For the exploratory meta-analyses, some studies reported multiple kappa-statistics for a given dementia type as they assessed more than one diagnostic criteria. Hence, we conducted separate analyses using the lowest (Figure S1) and highest (Figure S2) estimates. The results did not differ between them (0.01 difference). For the latter, we obtained a kappa-statistic of 0.72 (95% CI = [0.61, 0.82], 95% PI = [0.45, 0.98]) for the for presence or absence of all-cause dementia. For presence or absence of AD we obtained a kappa-statistic of 0.67 (95% CI = [0.58, 0.77], 95% PI = [0.42, 0.92]) and for presence or absence of VaD, we obtained a kappa-statistic of 0.80 (95% CI = [0.73, 0.87], 95% PI = [0.73, 0.87]).

3.4.3 | Sensitivity analyses

For the first main analysis using dichotomous classifications (Figure 2), the point estimates I-squared statistic suggests that the fraction of the between-study variance due to heterogeneity may be substantial (I-squared = 76) however the uncertainty is large (95% CI = [21, 97]). Those for the second and third analyses do not suggest
| Study                      | Subject No. | No. raters | Countries                          | Single site | Dementia types | Diagnostic criteria | Estimated prevalences | Study settings | Age information (in years) | %Female | Education information | Ethnicity | Estimated cognition/dementia severity information |
|---------------------------|-------------|------------|------------------------------------|-------------|----------------|---------------------|-----------------------|----------------|----------------------------|---------|-----------------------|-----------|---------------------------------------------------|
| Hogervost et al. (2003)   | 50          | 2          | UK                                 | Yes         | AD             | NINCDS-ADRDA, DSM-IV | 0.44 AD (based on post-mortem); 10/50 (0.20) | Secondary care and community | Mean = 78.4, SD = 8.6 (at death), range = 56-93 | -       | -                     | White     | -                   |
| Hogervost et al. (2000)   | 82          | 2          | UK                                 | Yes         | All-cause, AD  | NINCDS-ADRDA       | 0.91 all-cause, 0.72 AD (based on post-mortem) | Secondary care and community | Mean = 76.2, SD = 8.6 (at diagnosis), range = 54-91 | -       | -                     | White     | Mixed               |
| Chui et al. (2000)        | 25          | 7          | USA                                | No          | All-cause, AD, VaD | DSM-IV, NINCDS-ADHRA, NINDS-AIREN | 0.64 all-cause (14 mild, 3 moderate), 0.36 non-dementia (4 healthy, 4 MCI) | Secondary care | Mean = 67.8, SD = 9.1 (at study entry) | 48      | -                     | White     | More difficult to classify                       |
| Baldreschi et al. (1994)  | 51          | 14         | Canada, Chile, Malta, Nigeria, Spain, USA | No          | All-cause, AD, VaD | NINCDS-ADRDA, ICD-10, DSM-III-R | 0.70 all-cause (0.38 AD, 0.16 VaD, 0.16 other dementias, 0.30 non-dementia (0.20 MCI, 0.10 depression) | Secondary care | Mean = 64, range = 46-82 (at study entry) | 61      | -                     | Mixed     | Mixed               |
| Blacker et al. (1994)     | 60          | 2          | USA                                | No          | AD             | NINCDS-ADRDA       | 0.68 AD (based on consensus including autopsy) | Secondary care | Mean = 67.2, SD = 9.8 (at diagnosis) | 53      | Mean = 13.8 years, SD = 4.3 | White     | Easier to classify                                |
| Fratiglioni et al. (1992) | 668         | 2          | Sweden                             | Yes         | All-cause      | DSM-III-R           | 0.31 all-cause (estimated from mean of raters) | Community | Mean = 84.6, SD = 5 (at start of study) | 81      | 48% M & 36% FM had university education | White     | Easier to classify                                |
| Larson et al. (1996)      | 52, 29      | 4          | USA, Japan                         | No          | All-cause      | DSM-III-R           | 0.19 all-cause for 1995, 0.45 for 1996 | Community | Mean = 82.2, range = 67-98 | 64      | -                     | Mixed     | -                   |
| O Connor et al. (1996)    | 100         | 5          | Australia, Germany, Netherlands, USA, UK | No          | All-cause      | DSM-III-R           | 0.59 all-cause (estimated from mean of raters) | Secondary care and community | Mean = 80, range = 54-96 | 62      | -                     | White     | Mixed               |
| Study                  | Subject No. | No. raters | Countries | Single site | Dementia types | Diagnostic criteria | Estimated prevalences | Study settings | Age information (in years) | %Female | Education information | Ethnicity | Estimated cognition/dementia severity information |
|-----------------------|-------------|------------|-----------|-------------|----------------|---------------------|---------------------|----------------|--------------------------|---------|----------------------|-----------|--------------------------------------------------|
| Kay et al. (1998)     | 57          | 42         | UK        | No          | All-cause     | DSM-III-R          | 0.42 (estimated from ratings; 0.14 possible, 0.16 probable, 0.58 absent, 0.12 definite) | Community | Mean = 79.8, SD = 6.44, range = 68-93             | 67      | -                     | White     | Easier to classify                                              |
| Magaziner et al. (1996) | 100        | 2          | USA       | Yes         | All-cause     | DSM-III-R          | 0.54 all-cause (based on expert panel ratings) | Community | Mean = 81.3 (at admission to nursing home)         | 74      | -                     | White     | Easier to classify                                              |
| E. Graham et al. (1996) | 210        | 2          | Canada    | No          | All-cause, AD, VaD, NINCDS-ADRDA, DSM-III-R, ICD-10 | 0.39 all-cause (0.26 AD, 0.07 VaD, other dementias 0.06) | Community | 65 and over | -                        | -       | White                | -         |                                                    |
| Llamas-Velasco et al. (2016) | 30         | 27         | Spain     | Yes         | All-cause     | NIA-AA, DSM-III-R  | 0.60 all-cause/0.27 AD, 0.07 VaD, based on previous diagnoses | Secondary care | 55 and over | -                        | -       | White                | Mixed     |                                                    |
| LA Farrer et al. (1994) | 42          | 7          | USA, Canada | No          | AD            | NINCDS-ADRDA       | 0.69 AD (0.55 Pr AD, 0.14 Ps AD), 0.976 all-cause (1 patient non-dementia) | Secondary care | Mean = 70.2, SD = 8.7 (at onset) | 57      | Mean = 12.5 years, SD = 3.4 | White     | Mixed                                                   |
| IG McKeith et al. (1994) | 50          | 4          | UK        | Yes         | DLB           | McKeith Criteria   | 1.00 all-cause (0.40 DLB, 0.42 AD, 0.18 VaD) | Secondary care | Mean = 79.2, SD = 3.63 | 52      | -                     | White     | -                                                   |
| O.L. Lopez et al. (1990) | 40          | 4          | USA       | Yes         | AD            | NINCDS-ADRDA       | 0.75 AD (30 AD and 30 controls) | Secondary care | Mean = 67.1, SD = 8.54 | 48      | Mean = 4.2 years, SD = 2.02 | White     | Mixed                                                   |
| O.L. Lopez et al. (1994) | 42          | 4          | USA       | Yes         | VaD           | NINDS-AIREN        | 1.00 all-cause 0.34 (Pr + Ps VaD, avg of 4 raters) | Secondary care | Mean = 72.2, SD = 7.5 | 32      | Mean = 11.7 years, SD = 3.4 | White     | Easier to classify                                              |
| O.L. Lopez et al. (1999) | 40          | 4          | USA       | Yes         | AD, DLB, bvFTD | NINCDS-ADRDA, McKith, Lund and Manchester | 1.00 all-cause 0.25 AD, 0.20 DLB, 0.20 bvFTD, 0.20 PSP, 0.15 other | Secondary care | Mean = 67.8, SD = 7.31 (at start of study) | 48      | Mean = 13.2 years, SD = 3.9 | White     | Mixed (easier to classify for AD, more difficult for DLB, easier for bvFTD) |
| K LaMarre et al. (2013) | 20          | 18         | USA       | No          | bvFTD         | FTDC criteria      | 0.20 bvFTD (based on multidisciplinary consensus evaluation) | Secondary care | Median = 65, range = 54-85 (at start of study) | 30      | Median = 15 years (range = 12-20) | White     | -                                                   |
| Study                     | Subject No. | No. raters | Countries       | Single site | Dementia types | Diagnostic criteria | Estimated prevalences | Study settings       | Age information (in years) | %Female | Education information | Ethnicity | Estimated cognition/dementia severity information |
|--------------------------|-------------|------------|-----------------|-------------|----------------|---------------------|-----------------------|------------------------|--------------------------|---------|----------------------|-----------|--------------------------------------------------|
| Regier, narrow, Clarke (2013) et al. | 429         | 31 total (2 per patient) | USA, Canada | No | All-cause | DSM-5 | 0.13 all-cause (based on consensus DSM-5 MND) | Secondary care and community | Mean = 61.4, SD = 5.8 (at start of study) | 66 | 84% had greater than high school education | White | - |
| Gutzmann et al. (1990)   | 31          | 134 total (at least 2 per patient) | Germany, Austria, Switzerland | No | AD, VaD | ICD-10 | 0.03 AD, 0.03 VaD | Secondary care | - | - | - | White | - |
| Gutzmann et al. (1996)   | 39          | 451 total (at least 2 per patient) | Germany, Austria, Switzerland | No | AD, VaD | ICD-10 | 0.03 AD, 0.03 VaD | Secondary care | - | - | - | White | - |
| Sartorius et al. (1995)  | 3,493       | 942 total (at least 2 per patient) | 32 countries (UK, USA, others) | No | AD, VaD | ICD-10 | 0.01 AD, 0.01 VaD | Secondary care | 3.9% 70 and over, 6.7% 60-69, 10% 50-59, 79.4% 49 and under | 50 | - | Mixed | - |

Abbreviations: ADDTC, Alzheimer Disease Diagnostic and Treatment Centres; bvFTD, behavioural variant frontotemporal dementia; DSM, Diagnostic and Statistical Manual of Mental Disorders, Revised; FTDC, international behavioural variant TD criteria consortium criteria; ICD-10, International Classification of Diseases version 10; MND, Major Neurocognitive Disorder; NINCDS-ADRDA, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association; NINDS-AIREN, National Institute of Neurological Disorders and Stroke and Association Internationale pour la Recherche et l'Enseignement en Neurosciences; Pr = probable; Ps = possible; SD, Standard Deviation; VaD, vascular dementia

*Subject number refers to the number of subjects the study used for assessment of interrater agreement.

Refers specifically to the types of dementia for which agreement was assessed.

*N = 52 for 1995 analysis and N = 29 for 1996 analysis.
**Figure 2** Forest plot for interrater agreement, by dementia type & diagnostic criteria. Note: AD, Alzheimer’s dementia; DSM-III-R, Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition; ICD-10, International Classification of Diseases version 10; NINCDS-ADRDA, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s disease and Related Disorders Association; Pr, probable; Ps, possible; VaD, Vascular Dementia; Size of the squares corresponds to the study weight for Magaziner et al and Fratiglioni et al, we grouped together the ‘dementia’ and ‘intermediate’ categories into one ‘dementia’ category to calculate the kappa-statistic. Larson et al. conducted a second interrater agreement assessment after a standardisation meeting was held, and Fratiglioni et al conducted a second assessment which occurred after assessors saw reports from informants and interviewers case vignettes. We conducted the pooled analysis using the first set of analyses from these studies [Colour figure can be viewed at wileyonlinelibrary.com]

**Figure 3** Forest plot for interrater agreement for NINCDS-ADRDA AD, taking into account uncertainty of diagnosis. Note: AD, Alzheimer’s dementia; NINCDS-ADRDA, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s disease and Related Disorders Association; Pr, probable; Ps, possible; Size of the squares corresponds to the study weight [Colour figure can be viewed at wileyonlinelibrary.com]
that the fraction of the between-study variance due to heterogeneity is large (both equal to 0), however the uncertainty is large (95% CIs of [0.81] and [0.90] respectively), as well as those for the second meta-analysis which took uncertainty of diagnosis into account (Figures 3 and 4). The large degree of uncertainty in the I^2 statistic is not surprising, given the small number of studies. We did not perform meta-regression or subgroup analyses to explore sources of heterogeneity for the first meta-analysis, or any publication bias tests, due to the small number of studies. We performed leave-one-out analyses to investigate the influence of each study on the pooled estimates, and these did not change appreciably for either the main analyses (Figures S3, S4 and S5) or the exploratory analyses (Figure S6).

3.5 Narrative synthesis

In this section, we discuss some results (see Table 2) which were not pooled due to insufficient data. Results are quoted as X (Y, Z), where X is a point estimate, and (Y, Z) is a 95% confidence interval, or as ranges X-Y. Estimates are unweighted kappa-statistics unless otherwise stated.

3.5.1 All-cause dementia

One study observed good agreement for the NIA-AA (weighted kappa = 0.76 [0.65, 0.86]) amongst neurologists, assessing agreement for dementia versus MCI versus no cognitive impairment, and agreement did not vary substantially based on rater experience. This study also found kappa-statistics of 0.34 for discriminating mild from non-mild dementia, 0.44 for moderate versus non-moderate and 0.68 for severe versus non-severe, using the DSM-III-R severity classifications. For the DSM-5, a high-quality study found good agreement (0.78 [0.68, 0.88]) which was consistent across both centres [0.75 [0.59, 0.90] and 0.80 [0.65, 0.90]). The AC1 and AC2 statistics were able to calculate corresponded closely to the kappa-statistics or were slightly higher.

3.5.2 Alzheimer’s dementia

Two studies assessed agreement using the DSM-IV. One found agreement of 0.56 with large uncertainty (95% CI of [0.14, 0.98]) and the other 0.68 [0.48, 0.88]. Three studies assessed agreement using the ICD-10. One reported agreement (0.85, [0.75, 0.91]). Another found lower agreement (0.36 [0.19, 0.53]) and another reported k = 0.46 with no CI. One study looked at how estimated agreement varied by rater experience. Using three-way classifications (Pr vs. Ps vs. non-AD), They found that the two most experienced raters had the highest agreement (weighted kappa = 0.72 [0.56, 0.88]), however these two raters obtained amongst the lowest levels of agreement when using the dichotomous classification (weighted kappa = 0.63 [0.38, 0.89]). The AC1/2 statistics we were able to calculate corresponded closely to the kappa-statistics or were slightly higher.
| Study                      | Comparisons                                                                 | Kappa (95% CI)*                  | AC1 (95% CI)\(^{ab}\) |
|---------------------------|-----------------------------------------------------------------------------|----------------------------------|------------------------|
| All-cause dementia        |                                                                             |                                  |                        |
| DSM-IV                    |                                                                             |                                  |                        |
| Chui et al. (2000)\(^{19}\) | Dementia versus nondementia                                                 | 0.62 (0.27, 0.97)\(^{b}\)       | -                      |
| ICD-10                    |                                                                             |                                  |                        |
| Baldereschi et al. (1994)\(^{12}\) | Dementia versus nondementia                                                | 0.69 (0.40, 0.98)\(^{b}\)       | -                      |
| NINCDS-ADRDA              |                                                                             |                                  |                        |
| Chui et al. (2000)\(^{19}\) | Dementia versus nondementia                                                 | 0.65 (0.31, 0.99)\(^{b}\)       | -                      |
| Hogervorst et al. (2000)\(^{38}\) | Dementia versus nondementia                                                | 0.94 (0.81, 1)\(^{b}\)          | 0.98 (0.95, 1.00)      |
| NINDS-AIREN               |                                                                             |                                  |                        |
| Chui et al. (2000)\(^{19}\) | Dementia versus nondementia                                                 | 0.74 (0.45, 1)\(^{b}\)          | -                      |
| NIA-AA                    |                                                                             |                                  |                        |
| Llamas-Velasco et al. (2016)\(^{40}\) | Dementia versus MCI versus healthy (overall)   | 0.76 (0.65, 0.86) \(\text{[wtd.]}\) | -                      |
|                           | Dementia versus MCI versus healthy (senior neurologists)                    | 0.71 (0.59, 0.83) \(\text{[wtd.]}\) | -                      |
|                           | Dementia versus MCI versus healthy (junior neurologists)                    | 0.85 (0.73, 0.95) \(\text{[wtd.]}\) | -                      |
|                           | Dementia versus MCI versus healthy (neurology residents)                    | 0.69 (0.54, 0.81) \(\text{[wtd.]}\) | -                      |
| DSM-5                     |                                                                             |                                  |                        |
| Regier, narrow, Clarke et al. (2013)\(^{47}\) | Dementia versus nondementia | 0.78 (0.68, 0.88) | -                      |
|                           | Dementia versus nondementia (Mayo clinic)                                    | 0.75 (0.59, 0.90)                | -                      |
|                           | Dementia versus non dementia (UCLA)                                          | 0.80 (0.65, 0.90)                | -                      |
| DSM-III-R                 |                                                                             |                                  |                        |
| Baldereschi et al. (1994)\(^{12}\) | Dementia versus nondementia                                                | 0.67 (0.41, 0.93)\(^{b}\)       | -                      |

(Continues)
| Study                                      | Comparisons                                                                 | Kappa (95% CI)*          | AC1 (95% CI)\(^{AB}\) |
|--------------------------------------------|-----------------------------------------------------------------------------|--------------------------|------------------------|
| Fratiglioni et al. (1992)\(^{54}\)        | Dementia + uncertain versus nondementia (1\(^{st}\) comparison)            | 0.59 (0.51, 0.67)\(^{b}\) | 0.62 (0.56, 0.68)      |
|                                            | Dementia + questionable versus nondementia (2\(^{nd}\) comparison)        | 0.76 (0.71, 0.81)\(^{b}\) | 0.68 (0.80)            |
|                                            | Dementia versus nondementia versus uncertain (1\(^{st}\) comparison)      | 0.63 (0.57, 0.68) [wtd.]  | 0.75 (0.84)            |
|                                            | Dementia versus nondementia versus questionable (2\(^{nd}\) comparison)   | 0.75 (0.70, 0.80) [wtd.]  | 0.69 (0.64)            |
|                                            |                                                                            | [wtd.]                    | 0.74 [wtd.]            |
|                                            |                                                                            |                          | 0.81 (0.78, 0.85)      |
|                                            |                                                                            |                          | [wtd.]                  |
| Larson et al. (1998)\(^{55}\)             | Dementia versus nondementia (before standardisation meeting, N = 52)      | 0.44 (0.13, 0.75)\(^{b}\) | --                     |
|                                            |                                                                            | 0.90 (0.75, 1) \(^{b}\)  |                        |
|                                            | Dementia versus nondementia (after standardisation meeting, N = 29)       |                          |                        |
| O Connor et al. (1996)\(^{56}\)           | Dementia versus nondementia (mean within-centre)                           | 0.79 (−)                 | --                     |
|                                            | Dementia versus nondementia (mean between-centre)                          | 0.78 (−)                 |                        |
| Kay et al. (1998)\(^{57}\)                | Dementia versus nondementia (R stage\(^{d}\))                             | 0.69 (−)                 | ---                    |
|                                            | Dementia versus nondementia (V stage\(^{e}\))                             | 0.74 (−)                 |                        |
|                                            | Mild versus moderate versus severe (R stage\(^{d}\))                      | 0.22 (−) [unwtd.]         |                        |
|                                            | Mild versus moderate versus severe (V stage\(^{e}\))                      | 0.29 (−) [unwtd.]         |                        |
| Magaziner et al. (1996)\(^{58}\)          | Dementia + uncertain versus nondementia (expert panel vs. geriatrician)    | 0.66 (0.51, 0.81)\(^{b}\) | 0.66 (0.54, 0.79)      |
|                                            | Dementia versus nondementia versus uncertain (expert panel vs. geriatrician) | 0.66 (0.54, 0.79) [wtd.]  | 0.67 (0.54, 0.80)      |
|                                            | Dementia versus nondementia versus uncertain (panel neurologist vs. psychiatrist) | 0.66 (0.52, 0.80) [unwtd.] |                        |
|                                            |                                                                            |                          | [wtd.]:                 |
| Graham et al. (1996)\(^{59}\)             | Dementia versus nondementia                                               | 0.81 (0.73, 0.89)         | 0.84 (0.77, 0.91)      |
| Study                                      | Comparisons                                      | Kappa (95% CI) | AC1 (95% CI) |
|--------------------------------------------|--------------------------------------------------|----------------|--------------|
| Llamas-Velasco et al. (2016)\(^{40}\)      | Mild versus not mild (severity grading)           | 0.34 (−)       | --           |
|                                            | Moderate versus not moderate (severity grading)   | 0.44 (−)       | --           |
|                                            | Severe versus not severe (severity grading)      | 0.68 (−)       | --           |

**Alzheimer’s dementia (AD)**

| Study                                      | Comparisons                                      | Kappa (95% CI) | AC1 (95% CI) |
|--------------------------------------------|--------------------------------------------------|----------------|--------------|
| Hogervost et al. (2003)\(^{37}\)          | AD versus non-AD                                  | 0.68 (0.48, 0.88)\(^b\) | --           |
| Chui et al. (2000)\(^{19}\)               | AD versus non-AD                                  | 0.56 (0.14, 0.98)\(^b\) | --           |

**ICD-10**

| Study                                      | Comparisons                                      | Kappa (95% CI) | AC1 (95% CI) |
|--------------------------------------------|--------------------------------------------------|----------------|--------------|
| Gutzmann et al. (1990)\(^{48}\)           | AD versus non-AD                                  | 0.36 (0.19, 0.53)\(^b\) | --           |
| Gutzmann et al. (1996)\(^{50}\)           | AD versus non-AD                                  | 0.46 (−)        | --           |
| Sartorius et al. (1995)\(^{51}\)          | AD versus non-AD                                  | 0.83 (0.75, 0.91)\(^b\) | --           |

**NINCDS-ADRDA**

| Study                                      | Comparisons                                      | Kappa (95% CI) | AC1 (95% CI) |
|--------------------------------------------|--------------------------------------------------|----------------|--------------|
| Hogervost et al. (2000)\(^{38}\)          | Pr AD versus Ps AD versus non-AD                  | 0.74 (0.58, 0.90)\(^b\) | 0.84 (0.73, 0.95) |
|                                            | Pr AD versus Ps AD versus non-AD                  | 0.54 (0.39, 0.69)\(^{[wtd.]}\) | 0.60 (0.48, 0.72)\(^{[wtd.]}\) |
| Chui et al. (2000)\(^{19}\)               | Pr AD versus not Pr AD                            | 0.49 (0.05, 0.93)\(^b\) | --           |
|                                            | Ps AD versus Ps AD                               | 0.47 (0.13, 0.81)\(^b\) | --           |
|                                            | Ps AD versus not Ps AD                           | 0.46 (0.07, 0.85)\(^b\) | --           |
| Baldereschi et al. (1994)\(^{52}\)        | Pr AD versus not Pr AD                            | 0.58 (0.27, 0.89)\(^b\) | --           |
|                                            | Ps AD versus not Ps AD                           | 0.12 (−0.60, 0.84)\(^b\) | --           |
|                                            | Ps AD versus not Ps AD                           | 0.72 (0.49, 0.95)\(^b\) | --           |
| Blacker et al. (1994)\(^{53}\)            | Pr AD versus not Pr AD                            | 0.43 (0.20, 0.66)\(^b\) | 0.43 (0.20, 0.67) |
|                                            | Ps AD versus not Ps AD                           | 0.51 (0.29, 0.73)\(^b\) | 0.56 (0.35, 0.78) |
|                                            | Pr AD versus Ps AD                               | 0.46 (0.27, 0.66)\(^{[wtd.]}\) | 0.47 (0.27, 0.67)\(^{[wtd.]}\) |
| Graham et al. (1996)\(^{79}\)             | Ps AD versus not Ps AD                           | 0.76 (0.46, 0.86)\(^b\) | 0.76 (0.46, 0.86) |
| LA Farrer et al. (1994)\(^{61}\)          | Pr AD versus not Pr AD                           | 0.63 (0.38, 0.88) | --           |
|                                            | Pr AD versus not Ps AD                           | 0.59 (0.37, 0.81) | --           |
| Study                          | Comparisons                          | Kappa (95% CI) | AC1 (95% CI) |
|-------------------------------|---------------------------------------|----------------|--------------|
| O.L Lopez et al. (1990)³⁵     | Pr + Ps AD versus non-AD (overall)     | 0.72 (0.56, 0.88)³ | 0.76 (0.61, 0.92) |
|                               | Pr + Ps AD versus non-AD (rater 1 vs. 2) | 0.77 (0.56, 0.98) | 0.82 (0.64, 1) |
|                               | Pr + Ps AD versus non-AD (rater 1 vs. 3) | 0.74 (0.52, 0.95) | 0.76 (0.55, 0.97) |
|                               | Pr + Ps AD versus non-AD (rater 1 vs. 4) | 0.77 (0.56, 0.98) | 0.82 (0.64, 1) |
|                               | Pr + Ps AD versus non-AD (rater 2 vs. 3) | 0.65 (0.49, 0.82) | [wtd.] |
|                               | Pr + Ps AD versus non-AD (rater 2 vs. 4) | 0.57 (0.40, 0.75) | [wtd.] |
|                               | Pr + Ps AD versus non-AD (rater 3 vs. 4) | 0.54 (0.35, 0.74) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (overall) | 0.63 (0.46, 0.82) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (rater 1 vs. 2) | 0.72 (0.56, 0.88) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (rater 1 vs. 3) | 0.72 (0.56, 0.88) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (rater 1 vs. 4) | 0.57 (0.40, 0.75) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (rater 2 vs. 3) | 0.65 (0.49, 0.82) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (rater 2 vs. 4) | 0.57 (0.40, 0.75) | [wtd.] |
|                               | Pr AD versus Ps AD versus non-AD (rater 3 vs. 4) | 0.54 (0.35, 0.74) | [wtd.] |

| Study                          | Comparisons                          | Kappa (95% CI) | AC1 (95% CI) |
|-------------------------------|---------------------------------------|----------------|--------------|
| O.L Lopez et al. (1999)⁴⁵     | Pr + Ps AD versus non-AD               | 0.73 (0.57, 0.89)³ | 0.75 (0.60, 0.91) |
|                               | Pr versus Ps AD versus non-AD         | 0.68 (0.53, 0.83) [wtd.] | 0.73 (0.59, 0.88) [wtd.] |

Vascular dementia (VaD)

DSM-IV

Chui et al. (2000)⁴⁹

| Study                          | Comparisons                          | Kappa (95% CI) | AC1 (95% CI) |
|-------------------------------|---------------------------------------|----------------|--------------|
| Chui et al. (2000)⁴⁹           | VaD versus non-VaD                    | 0.59 (0.22, 0.96)³ | -             |
### TABLE 2 (Continued)

| Study                                      | Comparisons                        | Kappa (95% CI)* | AC1 (95% CI)ab |
|--------------------------------------------|-------------------------------------|-----------------|----------------|
| ICD-10                                     |                                     |                 |                |
| Baldereschi et al. (1994)                  | VaD versus non-VaD                  | 0.66 (0.29, 1)  | -              |
| Graham et al. (1996)                       | VaD versus non-VaD                  | 0.69 (0.48, 0.90)| 0.96 (0.93, 0.99) |
| Gutzmann et al. (1990)                     | VaD versus non-VaD                  | 0.80 (0.65, 0.95)| -              |
| Gutzmann et al. (1996)                     | VaD versus non-VaD                  | 0.49 (—)        | -              |
| Sartorius et al. (1995)                    | VaD versus non-VaD                  | 0.82 (0.70, 0.94)| -              |
| NINDS-AIREN                                |                                     |                 |                |
| Chui et al. (2000)                          | Pr versus not Pr VaD                | 0.42 (—0.48, 1) | —              |
| Chui et al. (2000)                          | Ps versus not Ps VaD                | 0.42 (—0.41, 1) | —              |
| O.L Lopez et al. (1994)                    | Ps + pr versus non-VaD              | 0.42 (—0.42, 1) | —              |
|                                             | Pr + Ps VaD versus non-VaD (overall)|                 | 0.86 (0.75, 0.97) |
|                                             | Pr + Ps VaD versus non-VaD (rater 1 vs. 2) | 0.94 (0.83, 1)  | 0.96 (0.88, 1)  |
|                                             | Pr + Ps VaD versus non-VaD (rater 1 vs. 3) | 0.74 (0.52, 0.95)| 0.78 (0.59, 0.98)|
|                                             | Pr + Ps VaD versus non-VaD (rater 1 vs. 4) | 0.84 (0.66, 1)  | 0.87 (0.73, 1)  |
|                                             | Pr + Ps VaD versus non-VaD (rater 2 vs. 3) | 0.79 (0.60, 0.99)| (0.65, 1)      |
|                                             | Pr + Ps VaD versus non-VaD (rater 2 vs. 4) | 0.89 (0.75, 1)  | 0.91 (0.79, 1)  |
|                                             | Pr + Ps VaD versus non-VaD (rater 3 vs. 4) | 0.80 (0.61, 0.99)| 0.91 (0.64, 1)  |
|                                             | Pr versus Ps VaD versus non-VaD (overall) | 0.84 (0.72, 0.96)| 0.87 (0.80, 0.96)|
|                                             | Pr versus Ps VaD versus non-VaD (rater 1 vs. 2) | 0.94 (0.83, 1)  | [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 1 vs. 3) | 0.74 (0.52, 0.95)| [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 1 vs. 4) | 0.84 (0.66, 1)  | [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 2 vs. 3) | 0.79 (0.60, 0.99)| [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 2 vs. 4) | 0.89 (0.75, 1)  | [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 3 vs. 4) | 0.80 (0.61, 0.99)| [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (overall) | 0.84 (0.72, 0.96)| [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 1 vs. 2) | 0.94 (0.83, 1)  | [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 1 vs. 3) | 0.74 (0.52, 0.95)| [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 1 vs. 4) | 0.84 (0.66, 1)  | [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 2 vs. 3) | 0.79 (0.60, 0.99)| [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 2 vs. 4) | 0.89 (0.75, 1)  | [wtd.]         |
|                                             | Pr versus Ps VaD versus non-VaD (rater 3 vs. 4) | 0.80 (0.61, 0.99)| [wtd.]         |

Dementia with Lewy bodies (DLB) (Continues)
| Study | Comparisons | Kappa (95% CI)* | AC1 (95% CI)ab |
|-------|-------------|---------------|---------------|
| McKeith Criteria (1992)<sup>16</sup> | DLB versus non-DLB (overall) | 0.64 (0.41, 0.93)<sup>b</sup> | --- |
| IG McKeith et al. (1994)<sup>42</sup> | DLB versus non-DLB (rater 1 vs. 2) | 0.63 (0.40, 0.86)<sup>b</sup> | --- |
| | DLB versus non-DLB (rater 1 vs. 3) | 0.52 (0.25, 0.79)<sup>b</sup> | --- |
| | DLB versus non-DLB (rater 1 vs. 4) | 0.50 (0.24, 0.76)<sup>b</sup> | --- |
| | DLB versus non-DLB (rater 2 vs. 3) | 0.69 (0.46, 0.92)<sup>b</sup> | --- |
| | DLB versus non-DLB (rater 2 vs. 4) | 0.60 (0.36, 0.84)<sup>b</sup> | --- |
| | DLB versus non-DLB (rater 3 vs. 4) | 0.52 (0.25, 0.79)<sup>b</sup> | --- |
| McKeith Criteria (1996)<sup>19</sup> | DLB versus non-DLB | 0.39 (0.10, 0.73) | 0.85 (0.74, 0.97) |
| O.L Lopez et al. (1999)<sup>45</sup> | FTD versus non-FTD | 0.75 (0.55, 0.97)<sup>b</sup> | 0.90 (0.81, 1) |
| Frontotemporal dementia (FTD) | | | |
| Lund-Manchester criteria | | | |
| O.L Lopez et al. (1999)<sup>45</sup> | Ps bvFTD versus non-Ps bvFTD | 0.81 (0.57, 1) | -- |
| | bvFTD vs. non bvFTD | 0.82 (0.66, 0.98) | -- |
| FTDC | K. LaMarre et al. (2013)<sup>46</sup> | | |
| | Ps bvFTD versus non-Ps bvFTD | 0.81 (0.57, 1) | -- |
| | bvFTD vs. non bvFTD | 0.82 (0.66, 0.98) | -- |

Abbreviations: AC1, Gwet’s AC1; bvFTD, behavioural variant FTD; CI, Confidence Interval; DSM, Diagnostic and Statistical Manual of Mental Disorders, Revised; FTDC, International Behavioural Variant FTD Consortium Criteria; NINCDS-ADRDA, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s disease and Related Disorders Association; ICD-10, International Classification of Diseases version 10; NINDS-AIREN, National Institute of Neurological Disorders and Stroke and Association Internationale pour la Recherche et l’Enseignement en Neurosciences; Pr, probable; Ps, possible; unwtd, unweighted; wtd, weighted

Note: Greyed out estimates indicate those included in the meta-analyses.

*Estimates not followed by square brackets are unweighted kappa; for estimates based on comparisons which are ordinal categories, [weighted] designates weighted kappa and [unweighted] designates unweighted kappa.

<sup>b</sup>Calculated from contingency table, observed agreement or standard error.

<sup>c</sup>Second comparison was held after refining some of the guidelines in the criteria and adding a questionable dementia category.

<sup>d</sup>Based on responses to questions from patient interviews and medical records.

<sup>e</sup>Based on responses to questions from patient interviews and medical records, informant interviews and vignettes written by lay interviewers.
3.5.3 | Vascular dementia

One study\(^{42}\) found very good agreement for the NINDS-AIREN distinguishing probable from possible Vascular dementia (VaD; AC2 = 0.87 [0.80, 0.96]), which was similar to that found in the meta-analysis for the presence or absence of VaD for any criteria (Figure 3). This study also investigated the effect of rater profession and found agreement for the presence or absence of VaD was highest between the two neurologists (AC2 = 0.96 [0.88, 1]), which was still the highest of the rater-pairs for distinguishing probable, possible and non-VaD. One study\(^{44}\) found a kappa-statistic of 0.59 using the DSM-IV (0.59), however the uncertainty was large (95% CI = [0.22, 0.95]). For one study,\(^{37}\) agreement was notably higher using AC1 compared to kappa.

3.5.4 | Dementia with Lewy bodies

Two studies assessed agreement for DLB using the McKeith criteria. One\(^{45}\) using the 1992 criteria found very good agreement (AC1 = 0.85 [0.74, 0.97]) which was underestimated using kappa-statistics. Another\(^{42}\) using the 1996 criteria also found good agreement (0.64 [0.41, 0.93]). The latter study also assessed agreement by rater-pairs and found highest agreement between the most experienced raters (0.88 [0.75, 1]).

3.5.5 | Frontotemporal dementia

One study\(^{46}\) assessing agreement for bvFTD using the FTDC criteria and found moderate-very good agreement for presence or absence of possible bvFTD (0.81 [0.57, 1.05]) and good-very good agreement for presence or absence of probable bvFTD (0.82 [0.66, 0.98]). They also did not find sufficient evidence to suggest that agreement was associated with rater experience\(^{46}\) (odds ratio = 1.03, 95% CI = [0.98, 1.08]). Another study\(^{45}\) found very good agreement for FTD using the Lund-Manchester criteria (AC1 = 0.90 [0.81, 1]).

4 | DISCUSSION

4.1 | Findings

The meta-analysis evaluating the DSM-III-R for all-cause dementia using five studies found 'moderate-good' agreement for presence or absence of dementia with notable between-study heterogeneity (95% PI = [0.40, 0.92]). For the NINCDS-ADRDA for AD using eight studies, we found 'good' agreement when probable and possible AD categories were combined, and for the ICD-10 we found 'good-excellent' agreement for presence or absence of VaD. For these dementia types and comparisons, we found similar results when pooling all studies together regardless of criteria used. For the NINCDS-ADRDA using 4 studies, when distinguishing probable and possible cases of AD we found that agreement dropped from 'good-very good' to 'moderate-good' using AC1/2. For other criteria for these dementia types, there was an insufficient number of studies to obtain a pooled estimate. We were able to compute Gwet’s AC1/2 coefficients for a subset of studies, and for dementia and AD compared to kappa most estimates increased slightly (<0.10), however some estimates for VaD, DLB and FTD increased substantially. This is in line with other studies\(^{39}\) which have found increased agreement when using AC1/2 compared to kappa-statistics, so estimates based on the latter may be underestimates of agreement. The estimates which used kappa-statistics may therefore be somewhat conservative. Only one study\(^{47}\) was rated as high quality. Few estimates obtained (both for kappa and AC1/2) were at the upper end (>0.90) of the 'very good' range.

A high-quality study\(^{47}\) found 'good-very good' agreement for the DSM-5 for presence or absence of dementia. A smaller, lower quality study\(^{48}\) suggested the NIA-AA diagnostic criteria has 'very good-good' agreement for distinguishing dementia MCI and normal cognition. For DLB, two studies using the McKeith criteria found 'good-very good' agreement, and for FTD two studies found 'very good' agreement. More studies should be conducted to further investigate these criteria. Findings were mixed for studies which looked at how agreement related with rater experience. Llamas-Velasco et al\(^{49}\) suggested that their study found slightly higher agreement for less experienced neurologists because they had recently received similar neurological training, whereas senior ones may have based diagnosis on their substantial and variable experience.

4.2 | Strengths and limitations

The main strength of this study is the fact that it is the first systematic-review and meta-analysis assessing interrater agreement in clinical dementia diagnosis. We used a comprehensive search strategy and assessed the most recent evidence on interrater agreement for dementia diagnosis, and computed multiple agreement coefficients where this was possible. This study has some limitations. Namely, we had to exclude studies not published in English language due to resource limitations. The estimates of agreement can vary from one study to another due to differences in study settings, rater and subject characteristics. Due to a limited number of studies that assessed agreement for any given criteria and dementia type, we did not have enough information to conduct subgroup analysis or meta-regression to investigate sources of heterogeneity, or to assess publication bias. We were also not able to compute AC1/2 statistics for all studies. Lastly, there was an insufficient number of studies to obtain pooled estimates for some criteria and dementia types.

4.3 | Comparison with other studies

An evidence-based review conducted in 2001\(^{60}\) included some of the studies that were part of our systematic review. This was not a
systematic review and they did not attempt meta-analysis. They concluded that DSM-III-R and NINCDS-ADRDA had good level of agreement for all-cause dementia and AD, which is similar to our conclusions for these criteria. Our study was a systematic review and meta-analysis, used a comprehensive search across 4 databases, and included more studies assessing agreement, including those released since the evidence-based review was published. Some previous studies have suggested that different diagnostic criteria may identify distinct groups of subjects. For example, one study investigated the effect of different diagnostic criteria on the prevalence of dementia, and they found that the diagnostic criteria affected the observed prevalence of dementia. In a sample of 1879 subjects, they found that 17.3% were diagnosed with dementia according to DSM-III-R, 13.7% for DSM-IV, and only 3.1% using ICD-10, and there was low agreement between the criteria. Conversely, another study found 100% agreement between DSM-IV and ICD-10 criteria. They suggested that the differing prevalence estimates seen in other studies between these criteria are due to the ICD-10 being interpreted so that people assume that decline in all of the executive functions must be fulfilled to justify a positive diagnosis, and they argue that the WHO should specify the criteria so that diagnosis is less open to interpretation. Agreement between criteria was outside the scope of this systematic review, however, this is an area which requires more research as it has potentially important clinical implications.

One study found similar levels of agreement to that found in this review for the diagnosis of coronary artery disease using CT angiography. Santelmann et al. carried out a systematic review and meta-analysis on various mental disorders. They pooled studies regardless of classification systems (e.g., iterations of DSM and ICD-10) used and found similar levels of agreement to that found in this review for Schizophrenia, Bipolar disorder and depression, but lower for schizoaffective disorder.

### 4.4 Implications for research and clinical practice

The fact that we found a relatively good level of agreement for the pooled measures is probably not surprising, considering that dementia has less room for subjectivity than other conditions that may be susceptible to more interrater disagreement. Since these tests are often assumed to be perfect in research, these findings should be thought of in the context of a perfect test. For example, even small levels of misclassification have potentially important implications for studies where the outcome is incident dementia. For instance, for RCTs, larger sample sizes may be needed to compensate as a result. One study investigated the impact of imperfect interrater agreement on trial power for stroke studies, and found that just 5% misclassification of both trial arms could result in a 20% increase in sample size required. This also has implications for cohort studies investigating the conversion rates from MCI to dementia.

The diagnostic criteria evaluated in this review are often used as reference tests and assumed to be 100% accurate in DTA studies in order to apply standard methods. Imperfect agreement has the potential to cause some non-negligible bias in the sensitivity and specificity of these reference tests, even if we assume the test itself is 100% accurate under some ‘perfect’ observer. For instance, in a low-prevalence population, even a ‘very good’ agreement (kappa = 0.8) could cause the sensitivity of the reference test to decrease from 100% to under 90%, which could influence the accuracy estimates of the tests under evaluation. Therefore, DTA research conducted on dementia tests should investigate the use of methods which can estimate accuracy without assuming a perfect reference test.

Another important area for future research is conducting further agreement studies for clinical dementia, particularly for the newer criteria and DLB and FTD. We suggest that future agreement studies do not only report kappa-statistics, which have a higher risk leading to misleading results and underestimating agreement due to the kappa paradox. Rather, they should report multiple agreement statistics including ones which may be more stable such as Gwet’s AC1/2, and convey appropriate uncertainty.

Fratiglioni et al. found improved agreement using the DSM-III-R after standardisation. Larson et al. also found that interrater agreement for all-cause dementia using the DSM-III-R improved from ‘moderate’ to ‘very good’ after a standardisation meeting was held, however the post-standardisation assessment was conducted on a subgroup of patients so this is likely an overestimate. Elisabeth C.W. et al. found slightly improved agreement for the radiological part of the NINDS-AIREN criteria for vascular dementia after standardising, which was notable in the experienced group (from ‘fair-moderate’ to ‘moderate-good’); however, no improvement was observed in the inexperienced group of raters. This suggests that it may be possible to further improve interrater agreement in clinical practice training and operationalisation of diagnostic criteria. Given the differing definitions across various diagnostic criteria, the process of standardisation would need to differ between the various diagnostic criteria for a given dementia type.

### 5 Conclusion

We found evidence to suggest that the DSM-III-R for dementia has ‘moderate-good’ agreement, NINCDS-ADRDA for AD ‘good’ for presence or absence of dementia and ‘moderate-good’ when differentiating probable and possible cases, and ‘good-very good’ using the ICD-10 for VaD, according to Altman classifications. Evidence was more limited for other criteria and dementia subtypes, and only one study was rated as high-quality, which suggested that the DSM-5 has ‘good’ agreement for presence or absence of dementia. We also found some evidence from a smaller study suggesting that the NIA-AA has ‘good’ agreement for distinguishing dementia, MCI and normal cognition. Few estimates obtained were at the upper-end of the ‘very good’ range (>0.90). Future research should verify these
findings, further investigate interrater agreement for clinical dementia (particularly for DLB and FTD, and for newer diagnostic criteria) using multiple measures of agreement, investigate the effect that relatively high (but imperfect) interrater agreement has on studies which assume clinical diagnostic criteria are perfect, and employ methods to adjust for imperfection.

ACKNOWLEDGEMENTS
Funding: EC is funded by a National Institute for Health Research Systematic Review Fellowship (project number RM-SR-2017-09-023). The Complex Reviews Support Unit is funded by the National Institute for Health Research (project number 14/178/29). The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health. The NIHR had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

CONFLICT OF INTEREST
The authors have no conflicts of interest to report.

DATA AVAILABILITY STATEMENT
Data available on request from the authors.

ORCID
Enzo Cerullo https://orcid.org/0000-0001-5036-4564

REFERENCES
1. Nelson PT, Braak H, Markesbery WR. Neuropathology and cognitive impairment in Alzheimer disease: a complex but coherent relationship. J Neuropathol Exp Neurol. 2009;68(1):1-14. https://doi.org/10.1097/NEN.0b013e3181919a48
2. Savva GM, Wharton SB, Ince PG, Forster G, Matthews FE, Brayne C. Age, neuropathology, and dementia published online. N Engl J Med. 2009;360:2302-2309. https://doi.org/10.1056/NEJMo0806142
3. Forlenza OV, Radanovic M, Talib LL, et al. Cerebrospinal fluid biomarkers in Alzheimer’s disease: diagnostic accuracy and prediction of dementia published online. Alzheimer’s Dementia Diagnosis, Assess Dis Monit. 2015;1:455-463. https://doi.org/10.1016/j.dadm.2015.09.003
4. Creavin ST, Wisniewski S, Noel-Storr AH, et al. Mini-mental state examination (MMSE) for the detection of dementia in clinically un-evaluated people aged 65 and over in community and primary care populations. Cochrane Database Syst Rev. 2016;1:CD011145. https://doi.org/10.1002/14651858.CD011145.pub2
5. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106. https://doi.org/10.1016/j.jclinepi.2010.03.002
6. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. 2009;62(10):e1-e34. https://doi.org/10.1016/j.jclinepi.2009.06.006
7. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research published online. Stat Med. 2002;21(14):2109-2129. https://doi.org/10.1002/sim.1180
8. World Health Organization (WHO). The ICD-10 Classification of Mental and Behavioural Disorders. 10th ed. Geneva: World Health Organization (WHO); 1993.
9. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-III-R). 3rd ed. Washington, DC: American Psychiatric Association; 1987.
10. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). 4th ed. Washington, DC: American Psychiatric Association; 1994.
11. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). 5th ed. Washington, DC: American Psychiatric Association; 2013.
12. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer’s disease: report of the NINCDS-ADRD A work group under the auspices of department of health and human services task force on Alzheimer’s disease. Neurology. 1984;34(7):939. https://doi.org/10.1212/wnl.34.7.939
13. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on aging-Alzheimer’s association workgroups on diagnostic guidelines for Alzheimer’s disease. Alzheimer’s Dement J Alzheimer’s Assoc. 2011;7(3):263-269. https://doi.org/10.1016/j.jalz.2011.03.005
14. Román GC, Tatemichi TK, Erkinjuntti T, et al. Vascular dementia: diagnostic criteria for research studies: report of the NINDS-AIREN international workshop. Neurology. 1993;43(2):250. https://doi.org/10.1212/wnl.43.2.250
15. Clinical and neuropathological criteria for frontotemporal dementia. The Lund and Manchester Groups. J Neurol Neurosurg Psychiatry. 1994;57(4):416-418. https://doi.org/10.1136/jnnp.57.4.416
16. Neary D, Snowden JS, Gustafson L, et al. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology. 1998;51:1546-1554. https://doi.org/10.1212/WNL.51.6.1546
17. Rascovsky K, Hodges JR, Knopman D, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain A J Neurol. 2011;134(Pt 9):2456-2477. https://doi.org/10.1093/brain/awr179
18. McKeith IG, Perry RH, Fairbairn AF, Jabeen S, Perry EK. Operational criteria for senile dementia of Lewy body type (SDLT) Published online. Psychol Med. 1992;22:911-922. https://doi.org/10.1017/S0033291700038484
19. McKeith IG, Galasko D, Kosaka K, et al. Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the consortium on DLB international workshop Published online. Neurology. 1996;47:1113-1124. https://doi.org/10.1212/WNL.47.5.1113
20. McKeith IG, Dickson DW, Lowe J, et al. Diagnosis and management of dementia with Lewy bodies: third report of the DLB consortium. Neurology. 2005;65(12):1863-1872. https://doi.org/10.1212/01.wnl.0000187889.17253.b1
21. McKeith IG, Boeve BF, Dickson DW, et al. Diagnosis and management of dementia with Lewy bodies: fourth consensus report of the DLB Consortium Published online. Neurology. 2017.
22. Higgins JPT Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. John Wiley & Sons; 2019.
23. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement Published online. Br J Math Stat Psychol. 2008;61:29-48. https://doi.org/10.1348/000711006X126600
24. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC; 2010.
neurology. Neurology. 2001;56(9):1143-1153. https://doi.org/10.1212/wnl.56.9.1143
61. Erkinjuntti T, Østbye T, Steenhuis R, Hachinski V. The effect of different diagnostic criteria on the prevalence of dementia. N Engl J Med. 1997;337(23):1667-1674.
62. Naik M, Nygaard HA. Diagnosing dementia—ICD-10 not so bad after all: a comparison between dementia criteria according to DSM-IV and ICD-10. Int J Geriatr Psychiatry. 2008;23(3):279-282.
63. Williams MC, Golay SK, Hunter A, et al. Observer variability in the assessment of CT coronary angiography and coronary artery calcium score: substudy of the Scottish computed tomography of the HEART (SCOT-HEART) trial Published online. Open Heart. 2015;2:e000234. https://doi.org/10.1136/openhrt-2014-000234
64. Santelmann H, Franklin J, Bultloff J, Baethge C. Interrater reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression—a systematic review and meta-analysis Published online. Schizophrenia Res. 2016;176:357-363. https://doi.org/10.1016/j.schres.2016.07.012
65. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia—meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand. 2009;119(4):252-265. https://doi.org/10.1111/j.1600-0447.2008.01326.x
66. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews Published online. J Clin Epidemiol. 2005;58:982-990. https://doi.org/10.1016/j.jclinepi.2005.02.022
67. Kunz, A. Misclassification and Kappa-Statistic: Theoretical Relationship and Consequences in Application. 2013.
68. Xu H, Black MA, Craig BA. Evaluating accuracy of diagnostic tests with intermediate results in the absence of a gold standard

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Cerullo E, Quinn TJ, McCleery J, Vounzoulaki E, Cooper NJ, Sutton AJ. Interrater agreement in dementia diagnosis: A systematic review and meta-analysis. Int J Geriat Psychiatry. 2021;36:1127–1147. https://doi.org/10.1002/gps.5499