HHS Public Access

Author manuscript

*Neuropsychol Rev*. Author manuscript; available in PMC 2018 April 05.

Published in final edited form as:

*Neuropsychol Rev*. 2017 December ; 27(4): 354–388. doi:10.1007/s11065-017-9360-6.

# Diagnostic Accuracy of Memory Measures in Alzheimer's Dementia and Mild Cognitive Impairment: a Systematic Review and Meta-Analysis

**Gali H. Weissberger**[1,2], **Jessica V. Strong**[2,3], **Kayla B. Stefanidis**[4], **Mathew J. Summers**[4], **Mark W. Bondi**[5,6], and **Nikki H. Stricker**[2,7]

[1]Brain, Behavior, and Aging Research Center, VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA

[2]Psychology Service, VA Boston Healthcare System, Boston, MA, USA

[3]New England Geriatric Research, Education and Clinical Center (GRECC), Boston VA Healthcare System, Boston, MA, USA

[4]Sunshine Coast Mind and Neuroscience – Thompson Institute, University of the Sunshine Coast, Sippy Downs, Queensland, Australia

[5]VA San Diego Healthcare System, San Diego, CA, USA

[6]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

[7]Department of Psychiatry and Psychology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA

## Abstract

With an increasing focus on biomarkers in dementia research, illustrating the role of neuropsychological assessment in detecting mild cognitive impairment (MCI) and Alzheimer's dementia (AD) is important. This systematic review and meta-analysis, conducted in accordance with PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) standards, summarizes the sensitivity and specificity of memory measures in individuals with MCI and AD. Both meta-analytic and qualitative examination of AD versus healthy control (HC) studies ($n$ = 47) revealed generally high sensitivity and specificity (≥ 80% for AD comparisons) for measures of immediate (sensitivity = 87%, specificity = 88%) and delayed memory (sensitivity = 89%, specificity = 89%), especially those involving word-list recall. Examination of MCI versus HC studies ($n$ = 38) revealed generally lower diagnostic accuracy for both immediate (sensitivity = 72%, specificity = 81%) and delayed memory (sensitivity = 75%, specificity = 81%). Measures that differentiated AD from other conditions ($n$ = 10 studies) yielded mixed results, with generally high sensitivity in the context of low or variable specificity. Results confirm that memory measures have high diagnostic accuracy for identification of AD, are promising but require further

✉Nikki H. Stricker, stricker.nikki@mayo.edu.

Gali H. Weissberger and Jessica V. Strong contributed equally to this work

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11065-017-9360-6) contains supplementary material, which is available to authorized users.

The authors report no conflicts of interest.

refinement for identification of MCI, and provide support for ongoing investigation of neuropsychological assessment as a cognitive biomarker of preclinical AD. Emphasizing diagnostic test accuracy statistics over null hypothesis testing in future studies will promote the ongoing use of neuropsychological tests as Alzheimer's disease research and clinical criteria increasingly rely upon cerebrospinal fluid (CSF) and neuroimaging biomarkers.

**Keywords**

Alzheimer'sdisease; Mild cognitive impairment; Neuropsychological testing; Memory; Sensitivity and specificity; Meta-analysis

## Introduction

Neuropsychological testing has demonstrated sensitivity to dementia, Mild Cognitive Impairment (MCI) and early preclinical stages of Alzheimer's disease (AD) and is relatively inexpensive. Only recently has neuropsychological testing been clearly listed as an important component of the diagnostic work-up for AD and MCI by the National Institute on Aging and Alzheimer's Association work groups (NIA-AA; MCI; Albert et al. 2011; McKhann et al. 2011) and for diagnosis of Major and Mild Neurocognitive Disorder (comparable to Dementia and MCI, respectively) in the DSM-5 (American Psychiatric Association 2013). However, despite the recognized importance and clear utility of neuropsychological testing (Bondi and Smith 2014), it is not a *required* component for diagnosis of AD or MCI in newly revised diagnostic systems (Albert et al. 2011; American Psychiatric Association 2013; McKhann et al. 2011). Just as is the case for neuropsychological testing, recent diagnostic systems have also incorporated biomarkers into updated consensus criteria for diagnosis of MCI due to AD and preclinical AD (Albert et al. 2011; McKhann et al. 2011; Sperling et al. 2011). However, the core clinical diagnostic criteria for AD do not include biomarkers, and biomarkers are seen only as "complimentary," serving to increase confidence that the clinical syndrome is due to the AD pathophysiological process (Jack et al. 2011). In addition, the use of biomarkers in preclinical AD and MCI are specifically prescribed for research and not for clinical purposes (Albert et al. 2011; Sperling et al. 2011). Nevertheless, biomarkers are often viewed as compelling additions to diagnosis and many clinical centers have adopted expensive and often invasive biomarker studies to aid in diagnosis of the AD pathological process, at times prior to ordering neuropsychological assessment. In addition, the the "A/T/N" (amyloid, tau, and neurodegeneration/neuronal injury) system is a recently proposed AD descriptive biomarker classification scheme (Jack et al. 2016), and it does not include cognition. However, recent evidence suggests that subtle cognitive decline alone can herald later development of biomarker positive states and mild cognitive impairment (MCI) or Alzheimer's dementia (Edmonds et al. 2015b), and cognitive differences are detectable in biomarker positive cognitively normal individuals (Han et al. 2017). One purpose of the present review and subsequent meta-analysis is to highlight the utility of neuropsychological testing as an equally valuable and arguably more affordable, less invasive cognitive biomarker of AD.

An illustration of how neuropsychological testing meets suggested guidelines for a useful biomarker may help to consolidate the evidence for the continued role of neuropsychology in the clinical diagnostic work-up. In the first review to do so, Fields et al. (2011) broadly outlined how neuropsychological testing may offer unique value as a biomarker for dementia. The current systematic review and meta-analysis further illustrates the utility of neuropsychology as a biomarker of AD by reviewing studies that report the diagnostic accuracy of memory measures in MCI and Alzheimer's dementia. To our knowledge, this is the first meta-analysis of the diagnostic accuracy of neuropsychological measures beyond cognitive screening measures.

A consensus report published in 1998 by The Ronald and Nancy Reagan Research Institute of the Alzheimer's Association and the National Institute on Aging Working Group on Molecular and Biochemical Markers of Alzheimer's Disease (Growdon et al. 1998; referred to as Consensus Workgroup hereafter) described the ideal features of a potential biomarker and operationalized criteria by which they can be evaluated. These criteria include range recommendations for sensitivity and specificity, most simply, that sensitivity and specificity should be no less than 80%within at least two independent studies distinguishing between patients with probable AD and normal control subjects. After this level of diagnostic accuracy is demonstrated, then further application in patients with possible AD or preclinical AD would be warranted. The consensus report also highlights that biomarkers can serve various purposes including diagnosis, screening, predicting conversion, monitoring disease progression, and detecting response to treatment. The value of any given biomarker may vary across its different applications. The more useful a biomarker is across settings, the higher its general value (see Fields et al. 2011 for a thorough discussion of how neuropsychological testing can serve most of these roles).

Although several reviews and meta-analyses have summarized diagnostic test accuracy statistics for the most commonly reported AD biomarkers, to our knowledge there has not been a review of diagnostic test accuracy statistics for neuropsychological measures. In fact, the relative lack of studies reporting diagnostic accuracy statistics was highlighted by Ivnik et al. (2000), who summarized this as a valid criticism of neuropsychology (reported in the 1996 Neuropsychological Assessment Panel of the American Academy of Neurology's Therapeutics and Technology Assessment Subcommittee). This gap in the literature was due mainly to an early over-reliance on null hypothesis testing and the unfortunate omission of diagnostic test accuracy statistics. The paucity of neuropsychological research and test manuals that include information about diagnostic validity is well recognized (Therapeutics and Technology Assessment Subcommittee of AAN 1996; Chelune 2010; Ivnik et al. 2000). This early overwhelming focus on null hypothesis testing has rendered much of the prior research demonstrating the utility of neuropsychology in assessment of dementia and MCI inapplicable at the individual clinical level. Fortunately, more studies recently have begun to include diagnostic test accuracy statistics, although these studies have yet to be summarized within one review.

The overall objective of this systematic review and meta-analysis was to evaluate the sensitivity and specificity of memory measures in individuals with MCI and AD. We hypothesized that the diagnostic accuracy of memory measures for studies comparing

individuals with AD and healthy controls (HC) would meet the minimum criteria put forth by the 1998 Consensus Workgroup.

## Method

This review was conducted in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (Moher et al. 2009). Although a review protocol was not registered prospectively, the primary objectives and methods were specified in advance. Meta-analyses were conducted whenever appropriate, and qualitative reviews were provided on measures of memory that were less widely reported or conceptually heterogeneous.

Articles to be considered for systematic review and meta-analysis were identified through a PubMed/MEDLINE search of studies that report diagnostic accuracy statistics for neuropsychological measures of memory for Mild Cognitive Impairment (MCI) or Alzheimer's disease (AD). Key words used for the search were [Neuropsychological Tests] or [Neuropsychology] and [Alzheimer] or [Mild Cognitive Impairment] and [sensitivity] or [specificity] or [ROC]. We identified and reviewed studies published before the date of our online database search on April 26, 2017 that included information regarding diagnostic accuracy of neuropsychological measures. Often, this was not the primary objective of the study. Additional studies were also identified through other sources including prior knowledge of studies, additional PubMed/MEDLINE searches outside the above search parameters, and review of references during screening. Because the vast majority of studies focused on memory, and to limit the scope of this review, studies were only included if they included an episodic memory measure. Some studies, particularly those with more complicated methodology or results, were reviewed by two or three reviewers (GW, JS, NS), whereas most studies were selected and reviewed by one reviewer (GW or JS). See Fig. 1 for a flow diagram describing the number of studies screened and meeting inclusion criteria.

Information on sensitivity and specificity was either directly extracted from the studies by the reviewers, or calculated using 2 × 2 tables that indicate number of false positives and negatives and true positives and negatives, if these data were presented. The extracted or calculated information is presented in the Online Resources (Tables i– iii). These tables present the author(s), year of publication, sample sizes for all groups, name of memory measure or neuropsychological test, sensitivity and specificity values, a cutoff value (if reported), whether the study used the test of interest in the diagnostic evaluation, and if the study reported 2 × 2 data. Studies were excluded if they did not report sensitivity and specificity data or sufficient information to calculate these statistics. We also excluded studies if 1) widely accepted diagnostic criteria for MCI or AD were not implemented (for AD - McKhann et al. 1984; McKhann et al. 2011; American Psychiatric Association 2000; for MCI - Albert et al. 2011; Petersen 2007; Petersen and Kanow 2001; Petersen 2004; Petersen et al. 2001; Petersen et al. 1999; Portet et al. 2006; Winblad et al. 2004), 2) sample characteristics could not be determined based on the information provided or the sample was heterogeneous (e.g., inclusion of comorbid neurological conditions in MCI or dementia samples, such as Parkinson's disease, that did not allow clear separation of results by suspected etiology), 3) a measure of episodic memory was not included, 4) methodology

appeared to be ambiguous or insufficiently specified (for example, vague reporting of specific neuropsychological measures used or unclear statistical analyses or results; in other words the neuropsychological measure used was unclear), or 5) published in a language other than English. In addition, studies investigating the diagnostic accuracy of screening measures were excluded as diagnostic validity of these measures for dementia have been previously reported (see Lin et al. 2013) and screening measures were viewed as beyond the scope of the current review. We only included studies that provided diagnostic accuracy statistics for scores reflecting memory, excluding several studies reporting combined scores of memory and other cognitive domains (e.g., naming and memory scores combined). In some cases, studies used the test under investigation as part of their diagnostic criteria used to classify their sample ("incorporation bias", see Noel-Storr et al. 2014). Although this circularity introduces bias and may overestimate the value of the diagnostic test (Noel-Storr et al. 2014), we chose to include these articles and identified them in the online resources (Tables i–iii). When feasible, we performed separate meta-analyses on studies of high and low quality to examine the potential influence of incorporation bias on results. Our criteria for "quality" was based on classifying each study for whether or not the measure of interest was used in the diagnosis ("Yes", "No", or "Unclear" if no explicit statement made by authors as to whether test was used in participant diagnosis).

Two types of cut-offs were typically used for the reported sensitivity and specificity. Optimal cutoffs are study-specific derivations that provide the best balance between sensitivity and specificity, typically derived through ROC analysis. Optimal cutoffs are typically represented as raw scores. Conventional cutoffs are based on an acceptable clinical standard (e.g., −1.5 standard deviations below the mean, etc.) derived from a test manual or other published normative data. If more than one conventional cut-off was reported in a study, we chose the value with the best balance of sensitivity and specificity to include here. Conventional cut-offs have the advantage of being more generalizable across studies and more easily applied to clinical settings, whereas optimal cutoffs maximize diagnostic accuracy regardless of whether the cut-off represents "impairment" in the clinical setting (e.g., many optimal cutoffs do not reach the minimally accepted clinical cut-off of more than or equal to −1 standard deviation, or SD, below the mean). Unfortunately, many studies did not report the specific value used as a cutoff. These studies were still included within the review and meta-analyses but are clearly noted in the online resources.

## Qualitative Review

Types of memory measures were divided into four categories to better evaluate patterns across studies: Immediate, Delayed, Associative Learning, and Other. "Immediate memory" was operationalized as recall of information directly following presentation of stimuli, such that measures using a distraction task or minutes of delay before recall were included under "delayed memory." Associative learning tasks require participants to bind together stimulus pairs (e.g., word pairs, object and location). This is distinguished from tasks that use higher-order category cues to assist in encoding or retrieval (e.g., selective reminding tasks or SRT). We included both verbal and visual associative learning tasks into this category, as well as measures of short-term visual memory binding (Parra et al. 2010; Parra et al. 2011). The "other" category was used for recognition memory, combined (e.g., immediate recall score

combined with yes/no recognition score) or interference scores, and other miscellaneous tests or indices. Interference (e.g., Fuld Object Memory Evaluation; Loewenstein et al. 2004) is any recall of an item or word that was not on the original list of presented stimuli. To optimally compare the diagnostic accuracy of different types of memory measures, we often report several different memory measures from the same study.

For both immediate and delayed categories, studies were further subdivided into verbal list free recall, verbal list cued recall or selective reminding, story free recall, visual free recall, retention (included in Delayed only), and "other." Free recall measures asked individuals to provide to-be-remembered information (e.g., a list, visual stimuli, or a story) from memory without any cues. Measures of cued-recall and combined measures of free- and cued-recall using a selective reminding paradigm[1] were included within the same division. Retention is the percent savings, or the number of words or items recalled on delay divided by the maximum number of words or items learned. Recognition tasks included yes/no recognition (e.g., "was car on the list you heard earlier?") and forced choice recognition (e.g., "Was car or banana on the list you heard earlier?"). Some studies included in the verbal memory section supplemented auditory stimuli with visual stimuli (e.g., pre-senting written words as they are read aloud or a corresponding picture with a to-be-learned word). Visual memory tasks included those with simple or complex geometric shapes, as well as route or map learning tasks.

For the qualitative review of studies, we used the minimum cutoff of 80%sensitivity and specificity for differentiating AD from HCs and from other dementias that was suggested by the Consensus Workgroup (1998). One limitation to our qualitative observations is that in general, direct comparisons across studies are confounded by varying methods and sample characteristics, prohibiting strong conclusions regarding which measures are most sensitive to AD. However, general patterns are discussed and where one published study directly examined two or more different memory measures, we comment on this comparison as appropriate. Studies also inconsistently reported cutoffs based on a conventional or an optimal cut point, further complicating direct comparison of diagnostic accuracy statistics across studies.

## Data Synthesis and Meta-Analysis

Meta-analyses were performed for immediate and delayed memory when an appropriate number of studies were available (minimum of 3 studies per test type was deemed as sufficient). It is important to note that for meta-analyses performed here using a single dependent variable (rather than several dependent variables considered jointly), where there are less than 5 studies there are important limitations to the validity of the meta-analysis, at k = 3 it is not possible to compute a rho correlation coefficient to assess potential threshold effects. Some studies listed in Tables i and iii were not included in meta-analysis due to

[1]The selective reminding paradigm has been well described (see Carlesimo et al. 2011) and is best exemplified by the Free and Cued Selective Reminding Test (Grober and Buschke 1987). This paradigm attempts to control for encoding of material by providing the name of each semantic category (4 total) and asking subjects to point to each of 4 items within the semantic category. This is followed by immediate category cued recall (repeated until 4/4 items recalled). Next, individuals are asked to freely recall the items on the list. A category cued recall procedure is then used for any items not freely recalled. This procedure is typically repeated three times and provides a measure of free recall and total recall (free recall + cued recall).

concern about duplication of subjects. For example, it is probable that a significant proportion of the data report by Chapman et al. (2016) is already represented by other studies as it is drawn from the National Alzheimer's Coordinating Centre (NACC database). Associative Learning and Other categories were not included in meta-analyses due to heterogeneity of measures. All meta-analyses were performed using R package 'Mada,' designed specifically for meta-analysis of diagnosticity data (Doebler 2015; Doebler and Holling 2012; Schwarzer et al. 2015). Using specificity and sensitivity values for each test, contingency data for each study (true positives, false positives, false negatives and true negatives) were computed using Microsoft Excel. Contingency data was rounded to whole numbers (≥0.5 rounded up, <0.5 rounded down). Contingency data and *k* were then entered into R to perform meta-analyses.

**Univariate Analysis—**Equality of sensitivity and specificity proportions was examined by $\chi^2$ test. The sensitivity and specificity values depend on the cut-off values used by different studies. Lowering the cut-off improves sensitivity but reduces specificity, whereas increasing the cut-off reduces sensitivity but increases specificity. The relationship between sensitivity and specificity as determined by the cut-off threshold is important to consider when performing meta-analyses of diagnostic test accuracy data where cut-off thresholds are likely to vary between studies included in the meta-analysis. Threshold effects were examined by Spearman rho correlation (sensitivity and false positive rate (1 – specificity)), with correlations ≥0.6 indicating potential threshold effects. The correlations are usually in a positive direction, but can be negative in direction. Diagnostic Odds Ratios (DOR) were calculated using the DSL method (DerSimonian and Laird random-effects; DerSimonian and Laird 1986). Coupled forest plots were used to examine threshold effects. Forest plots displaying an inverse relation (V or an inverted V pattern) indicate potential threshold effects. Where threshold effects are identified, interpretation of analyses should be based on descriptive analyses. Heterogeneity can be identified when the probability of the Q *statistic falls below*. 10. However, this statistical criterion may be less appropriate for meta-analysis of diagnostic tests which employ bivariate outcomes (sensitivity and specificity; Kim et al. 2015; Lee et al. 2015). Tau-squared quantifies the variance across studies, with a value of zero indicating minimal or no heterogeneity in the data.

**Hierarchical Meta-Analyses—**Following methods outlined by Kim et al. (2015) and Lee et al. (2015), we employed hierarchical methods, known as the bivariate and Rutter & Gatsonis hierarchical summary receiver operating characteristic (HSROC) models, respectively. These are random effect models in that they account for variance within studies as well as across studies. The use of such hierarchical methods is recommended (Lee et al. 2015) because these methods also account for the relationship between sensitivity and specificity, thereby directly addressing potential threshold effects. These methods produce the same results when no covariates are considered. Bivariate random-effects model, restricted maximum likelihood (REML) estimation, was employed with continuity correction set at 0.5. Some studies examined reported sensitivity or specificity values of 100. As such, the contingency data included values of 0. Such values have been noted to undermine the statistical validity of the meta-analysis. We addressed this by adding a small

continuity correction of 0.5 (default option) to each study, where required (Doebler and Holling 2012).

Studies were excluded if sensitivity and specificity values were missing or where 2 × 2 contingency data were missing, including those studies reporting a "set" cut-off value. For main analyses (AD Immediate, AD Delayed, MCI Immediate, MCI Delayed) single studies reporting multiple data points were statistically combined to form a single synthetic score. Synthetic scores were computed using the hierarchical methods described above, irrespective of sample size. As the combination of different types of measures of immediate and delayed recall into single meta-analyses may create additional variability, we conducted a series of subsequent meta-analyses of specific subclasses of immediate recall (list free recall, list cued selective reminding, story free recall, visual free recall) and subclasses of delayed recall (list free recall, list cued selective reminding, list retention, story free recall, visual free recall). For subclass analyses where single studies reported multiple data points, a single data point from each study was selected rather than calculation of synthetic score. The method used to select the single representative data point in these cases was to select the same measure as was used in other studies contained in the meta-analysis, and where this was not possible to identify the measure with the closest construct similarity to the other measures contained in the meta-analysis.

## Results

Descriptive (univariate) data and pooled estimates for AD and MCI can be found in Tables 1, 2, 3, 4, 5 and 6 respectively. Overall, the series of meta-analyses indicate that while immediate and delayed memory measures have high diagnostic accuracy in identifying AD, their capacity to discriminate between MCI and healthy persons is adequate but lower. For all analyses, the SROC (summary receiver operative characteristic) presented plots sensitivity values against false positive rate (FPR = 1 – specificity). Careful inspection of the SROC curves for MCI indicate substantial heterogeneity in sensitivity and specificity values across studies. We review quantitative and qualitative results for each subgroup of analyses.

### Alzheimer's Disease Versus Healthy Controls

We found a total of 84 studies comparing AD and HC based on our literature review and PubMed search criteria described above. After more careful review, 37 studies were excluded per the exclusion criteria described in the Method. We included 47 total studies for AD versus HC, many of which provide the sensitivity and specificity for multiple measures. Of the 47 studies, four studies explicitly stated that the measure of interest (in combination with other measures and clinical information) was considered during diagnosis, and in nine additional studies, this could not be determined based on the method sections. Eleven of the studies did not report the cut-off used or derived for the sensitivity and specificity values. Only four studies provided 2 × 2 data in the article (Cahn et al. 1995; O'Connell et al. 2004; Parra et al. 2010; Welsh et al. 1991). Almost all used diagnostic criteria of the National Institute of Neurological and Communicative Disorders and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA). See online resources (Tables i and ii) for additional information about diagnostic criteria applied by each study. In summary, over

half (54%) of the AD studies (included in both AD versus HC and AD versus Other; k = 29 of 54) included “probable” AD diagnoses only (i.e., excluded “possible AD” participants), 23%(k = 12) included both probable and possible AD diagnoses, k = 2 included confirmatory autopsy data (Salmon et al. 2002; Storandt and Morris 2010), one used familial gene sequence (Parra et al. 2010), and 19% (k = 10) did not specify probable or possible AD diagnoses.

**Immediate Memory—**Twenty-six data points contributed to meta-analysis of immediate recall measures in differentiating AD from HC (Table 1 and Fig. 2). Overall, these measures demonstrated excellent diagnostic accuracy with values well exceeding the suggested minimum cut-off values with a 95% confidence interval (95% CI) for sensitivity (Se = .87, 95% CI [.83, .90]) and specificity (Sp = .88, 95% CI [.85, .90]). Visual inspection of the forest plots (Fig. 2) and SROC curves (Fig. 3) supports this conclusion.

Due to the potential bias in some studies where the measures examined were also used to diagnose participants with AD, we classified all 26 papers according to whether or not the measure was used to diagnose participants. A total of 18 studies did not use the measure under examination for diagnosis of participants, with 3 studies using the measure for diagnosis and the remaining 5 studies being unclear as to how the measure was used in participant diagnosis. A meta-analysis of the 18 studies not using the measure for diagnosis was conducted (Table 2), indicating that the immediate memory recall measures continued to display excellent diagnostic accuracy for differentiating AD from HC with values well exceeding the suggested minimum cut-off for sensitivity (Se = .86, 95% CI [.82, .90]) and specificity (Sp = .88, 95% CI [.84, .92]). Visual inspection of the forest plots and SROC curves (Supplemental Figures xvii and xviii) supports this conclusion. However, the rho correlation did exceed the cutoff for potential threshold effects.

Forest plots for subclasses of immediate recall measure are presented in Online Resources (Figures i, ii, iii, and iv). Immediate memory recall subclasses (Table 1) generally displayed good to excellent sensitivity and specificity - List Free Recall Se = .87, 95%CI [.83, .90], Sp = .88, 95%CI [.85, .91]; List Cued Selective Reminding Se = .87, 95% CI [.78, .93], Sp = .93, 95% CI [.87, .97]; Visual Free Recall Se = .92, 95% CI [.86, .96], FPR = .90, 95% CI [.78, .95], with the exception of Immediate Story Recall which showed adequate sensitivity and specificity (Story Free Recall Se = .71, 95% CI [.61, .78], Sp = .75, 95% CI [.58, .86]). Visual inspection of the SROC curves for Immediate List Free Recall, Immediate List Cued Selective Reminding, and Visual Free Recall tests (Fig. 4) confirms that these measures display good diagnostic accuracy for differentiating AD from healthy controls. Story Free Recall displays lower diagnostic accuracy, however, it is important to note that due to small numbers of studies caution must be exercised in drawing firm conclusions regarding the diagnostic accuracy of Story Free Recall (k = 3) and Visual Free Recall measures (k = 4). Although no concerns were raised in the inspection of forest plots, correlations between sensitivity and FPR for AD Immediate List Cued or Selective Reminding and AD Immediate Visual Free Recall (cannot be calculated for Story Free Recall due to k = 3 cases) exceeded the cut-off for potential threshold effects (≥.60).

Immediate memory indices or factor scores were reported in two studies but were not included in the meta-analysis due to insufficient data (<3 studies; see Online Resources - Table i). Both the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) Immediate Memory Index and Mayo Cognitive Factor Scales (MCFS) Learning Recall Factor Score demonstrated higher than minimum cutoff for both sensitivity and specificity. However, for both of these studies (Duff et al. 2008; Ivnik et al. 2000), the test of interest was available, in addition to multiple other measures, when determining the diagnostic status of study participants, which could lead to inflated diagnostic accuracy statistics.

In summary, immediate memory measures, including immediate list free recall, immediate list cued or selective reminding, and immediate visual free recall demonstrated high diagnostic accuracy for differentiating AD from HC, with values well exceeding the suggested minimum cut-off for sensitivity and specificity (> .80). Immediate story recall measures displayed lower diagnostic accuracy. However, very few studies using story recall measures could be incorporated into meta-analysis.

**Delayed Memory—**Twenty-seven data points contributed to the meta-analytic evaluation of delayed recall measures in differentiating AD from HC (Table 3 and Fig. 5). Overall, these measures demonstrated excellent diagnostic accuracy with values well exceeding the suggested minimum cut-off for sensitivity (Se = .89, 95% CI [.87, .91]) and specificity (Sp = .89, 95% CI [.87, .91]). Visual inspection of the forest plots (Fig. 5) and SROC curves (Fig. 6) supports this conclusion.

To assess for potential bias in some studies where the measures examined were also used to diagnose participants with AD, we classified studies according to whether or not the measure was used to diagnose participants. A total of 18 (two data points from Delgado et al. 2016- Delgado 2016a, b) studies did not use the measure under examination for diagnosis of participants, with two studies using the measure for diagnosis and the remaining six studies being unclear as to how the measure was used in participant diagnosis. A meta-analysis of the 19 studies not using the measure for diagnosis was conducted (Table 2), indicating that the delayed memory recall measures continued to display excellent diagnostic accuracy for differentiating AD from HC with values well exceeding the suggested minimum cut-off for sensitivity (Se = .89, 95% CI [.85, .91]) and specificity (Sp = .89, 95% CI [.86, .91]). Visual inspection of the forest plots and SROC curves (Supplemental Figures xix and xx) supports this conclusion.

Forest plots for subclass analyses are presented in Online Resources (Figures v, vi, vii, viii, and ix respectively). Delayed memory subclasses displayed good to excellent sensitivity and specificity - List Free Recall Se = .90, 95%CI [.86, .92], Sp = .87, 95% CI [.84, .89]; List Cued Selective Reminding Se = .91, 95% CI [.87, .94], Sp = .92, 95% CI [.88, .95]; List Retention Se = .84, 95%CI [.73, .91], Sp = .81, 95% CI [.77, .84]; Visual Free Recall Se = .86, 95% CI [.82, .89], Sp = .88, 95% CI [.85, .91]; Story Free Recall Se = .93, 95% CI [.84, .98], Sp = .89, 95% CI [.79, .94]). Visual inspection of the SROC curves for the five subclasses of delayed memory recall (Fig. 7) confirms that all measures display good diagnostic accuracy for differentiating AD and HC. However caution is warranted in

interpreting data for List Cued Selective Reminding and Story Free Recall, as these exceeded the cut-off for potential threshold effects.

A few data points could not be included in the meta-analysis due to insufficient data (<3 studies), including story retention, visual retention, and other delayed scores that combined multiple indices. Values for story and visual retention were variable based on qualitative review. One study of story percent retention (WMS-R, 1987, Logical Memory percent retention; Testa et al. 2004) did not have sensitivity values that met the suggested minimum cutoff (Consensus Workgroup 1998), whereas another study (Clark et al. 2010) demonstrated values for RBANS-Story Retention into the 90s for specificity and sensitivity. Two studies reported values for visual retention and only one value, specificity of WMS-R Visual Reproduction Savings (Cahn et al. 1995), was above the minimum suggested cutoff (Consensus Workgroup 1998). The Mayo Cognitive Factor Score (MCFS) Retention score (Ivnik et al. 2000), derived from Wechsler Adult Intelligence Scales –Revised (WAIS-R), WMS-R, and Rey Auditory Verbal Learning Test (RAVLT), performed better than any individual retention indices. Duff et al. (2008) reported sensitivity and specificity values for the RBANS Delayed Memory Index, an index incorporating delayed word list recall, story recall, recognition, and delayed visual recall. The index had high sensitivity and specificity (92%) at a − 1.5 SD conventional cutoff.

Few studies provided data allowing for qualitative comparison of diagnostic accuracy across different memory measure types within the same study. Baek et al. (2012) reported values for immediate, delayed, and recognition memory on list learning (Korean Hopkins Verbal Learning Test; K-HVLT) and story learning (Korean Story Recall Test), with results suggesting list learning confers higher diagnostic accuracy relative to story recall, particularly for immediate recall trials. The diagnostic accuracy of recognition was poorer than recall for both stories and the word list. Duff et al. (2008) compared subtests of the RBANS in differentiating between AD and HCs. RBANS List Learning demonstrated good sensitivity and specificity using a conventional −1 SD cutoff and showed better balance across both sensitivity and specificity when compared to the RBANS Story Memory (immediate) in this study, although both measure types had excellent sensitivity and specificity after a delay. Salmon et al. (2002) reported higher sensitivity (98%) for delayed list recall (CVLT) relative to delayed story recall (WMS Logical Memory, 87% sensitivity), although specificity values were similar. Parra et al. (2012) reported immediate and delayed recall for a word list task, with values favoring immediate recall (sensitivity = 90%, specificity = 80%) over delay (sensitivity and specificity = 80%). Finally, Park et al. (2016) reported immediate and delayed cued recall on the RI-24 task. The delayed task showed higher sensitivity (89%) compared to the immediate task (75%) and the specificity was equivalent (91%). Fourteen studies presented data for both immediate and delayed verbal memory tasks (including both list and story). When compared directly within the same sample, most studies did not show large differences between immediate and delayed memory tasks for story or list-learning. Five studies showed a small improvement for either sensitivity or specificity on the delayed memory tasks, suggesting a possible advantage in diagnostic accuracy for the delayed task (e.g., Gavett et al. 2009). However, the remaining nine studies showed no such difference and some even demonstrated the opposite pattern –

stronger diagnostic accuracy values on immediate memory, compared to the delayed memory (e.g., Bertolucci et al. 2001).

In summary, delayed verbal memory tests of free and cued list and story memory, as well as non-verbal or visual tasks demonstrated good to excellent sensitivity and specificity for differentiating AD patients from HC participants. Delayed free recall of word list and delayed free recall of stories both consistently demonstrated strong sensitivity and specificity values above the minimum suggested cutoff (Consensus Workgroup 1998). Percent retention tended to have lower sensitivity and specificity. It is important to note that savings or retention is dependent on initial encoding, and therefore the sensitivity and specificity values may be artificially lowered. For example, if a patient only learns one item and remembers that one item, the calculated score will be 100% retention. The specificity of percent retention may be additionally important when comparing AD with other disorders (e.g., Vascular Dementia, Huntington's, or Parkinson's disease; Lundervold et al. 1994). Results suggest that immediate and delayed memory tasks may be similar in their diagnostic accuracy (meta-analysis findings show.86 and .88 sensitivity, .89 and .89 specificity, respectively) for differentiating AD patients from HC participants. In a clinical context, immediate memory tasks require much less time compared to delay tasks for both patients and examiners, thus it is important to determine whether delayed measures offer superior diagnostic accuracy relative to immediate memory.

**Associative Learning—**Seven studies, two of which used the same sample (Parra et al. 2010, 2011) were included, reporting data for ten tasks. The Visual Association Test (VAT)[2] had sensitivity (83%) and specificity (91%) values above the suggested cutoff (Lindeboom et al. 2002). A second paradigm[3] used by Lowndes et al. (2008) measured performance on Verbal Paired Associate-Recognition and Verbal Paired Cued-Recall tasks. Both of these tasks demonstrated strong specificity (100 and 96, respectively) and sensitivity (86%), meeting the recommended cutoff of 80% (Consensus Workgroup 1998). Storandt and Morris (2010) reported lower than expected values for both sensitivity (62%) and specificity (70%) on the WMS Associate Learning Immediate Recall task. Though they reported multiple cutoff values in the article, a standard deviation of −0.5 was the best balance between sensitivity and specificity. O'Connell and colleagues (O'Connell et al. 2004) reported 100% specificity and only 68% sensitivity on the Cambridge Neuropsychological Test Automated Battery (CANTAB) – Paired Associates Learning (PAL) test,[4] suggesting that the test may not meet the minimum criteria for detecting AD. In a comparison of the paper and pencil and computerized versions of The Placing Test,[5] diagnostic accuracy was equivalent

[2]A picture with two interacting pieces is presented (e.g., ape holding an umbrella) and individuals are asked to name both aspects of the picture. Immediately following presentation of six pictures, six cards are presented containing only one aspect of the picture (e.g., the ape alone). With presentation of the card missing an object, the individual is asked to say the item that is missing. Responses are accepted in any format: written, drawn, oral, or mimed.

[3]Eight semantically or associatively unrelated word pairs (four concrete, e.g., horse-forest, and four abstract, e.g., open-fresh). Associate-Recognition task was always completed during the first and the Cued-Recall during the second session. After presentation of the eight pairs, in the Associate-Recognition phase, a cue was presented at the top of the page and individuals identified which from a list of 4 items had been presented with the cue. For the Cued- Recall phase, the cue was presented and the individual was asked to recall the word that had been paired with the cue.

[4]A computerized task that displays boxes on the screen that are randomly opened. Some boxes contain a pattern. Patterns are then displayed in the middle of the screen and the individual must match the pattern in the middle of the screen to the appropriate location (i.e., box). When a mistake is made, the boxes are re-opened to remind the individual. The test becomes more difficult throughout and takes about 10 min to administer.

(Vacante et al. 2013). Specificity for both was reportedly 79%, which is just below the suggested minimum cutoff. The total score of the computerized test, which included faces, objects, and an additional 10 items, was reported to have equal sensitivity to the others (89%) and improved specificity (93%).

Parra et al. (2010) reported adequate sensitivity (82%) and specificity (77%) on a traditional associative learning task[6] in individuals with early-onset familial AD. Parra and colleagues also created a novel "visual short term memory binding" task[7] based on a change detection paradigm. The sensitivity (77%) and specificity (83%) for this test were also adequate. Of note, however, is that their study also included asymptomatic individuals who were known carriers of the E280A mutation who did not meet criteria for AD or MCI. Sensitivity of the binding condition in these individuals was 73%, a promising value given that they are in the preclinical phase of AD. For comparison, sensitivity of the WMS Verbal Paired Associates (VPA) was 40% for asymptomatic carriers. Another study published by Parra et al. (2011) found specific deficits in color-color short-term memory binding in both sporadic and familial AD. Sporadic AD cases demonstrated a 79% specificity and sensitivity, whereas familial AD cases demonstrated 77% sensitivity and 100% specificity for the bound colors condition.

Overall the associative learning tasks varied in terms of their sensitivity and specificity values. Of the ten measures reported, two-thirds demonstrated specificity values that were above the minimum cutoff. Similarly, seven of the ten measures demonstrated sensitivity values that were above the minimum cutoff. However, all but one study (Lindeboom et al. 2002) had small sample sizes ranging from 18 to 55, resulting in large confidence intervals.

**Other Memory Measures—**Ten studies that included recognition memory studies reported data on fourteen recognition tasks (yes-no recognition and forced-choice paradigms). All but one (Consortium to Establish a Registry for Alzheimer's Disease - CERAD -Word List Recognition – Finnish version; (Sotaniemi et al. 2012) reported specificity values above the minimum suggested cutoff of 80% (Consensus Workgroup 1998). In contrast, only 5 of the 14 sensitivity values reported met the minimum suggested cutoff. This pattern suggests that if recognition memory is impaired, it is likely to indicate AD (high specificity), but if it is not impaired we cannot be confident about ruling out AD (low sensitivity). Future research needs to examine other recognition memory tasks that may

[5]In the paper and pencil task, 10 faces and 10 objects were presented. The original version of the test showed faces in black and white whereas the novel version presented faces and objects in color. After encoding, individuals are immediately presented with the face or object (one at a time) and report in which quadrant of the page the item was originally presented. The total score is the number placed correctly out of 20. The computerized version was similar with an additional 10 shapes or 10 animals (30 items presented total) and participants were asked to click on the quadrant where the item had initially been presented. For both versions of the test, the 10 items belonging to a category were presented and tested separately, for example, encoding and testing phase of 10 faces followed by encoding and testing of 10 objects.

[6]WMS Verbal Paired Associates (Wechsler 1945, VPA) (Spanish translation), individuals learned 10 pairs of words (6 related, 4 unrelated) across three trials, providing one score of total memory acquisition.

[7]Individuals saw 2 or 3 items on the screen (difficulty varied depending on diagnosis; HC saw 3 items, AD patients saw 2 items) for 2000 ms during the study display. Following the study phase, individuals' visual short-term memory was assessed for single feature and binding by using "color only" and "shape only" conditions in addition to "shape-color binding" conditions. In single feature conditions, new shapes or new colors were replaced in the test phase, respectively, so that memory for individual features was required to detect the change. In the shape-color binding condition, two shapes changed colors in the test phase compared to the study phase so that memory of both the bound shape and color elements was required to detect change.

have better overall diagnostic accuracy (e.g., California Verbal Learning Test-II Total Recognition Discriminability).

All of the studies reviewed that were included under the Combined or Interference category (k = 8) reported sensitivity values well above the minimum recommended cutoff (Consensus Workgroup 1998) at 85% or higher. In addition, all of the specificity values were excellent, with all but one above 90%. It is difficult to draw conclusions on this category as a whole due to the variability of types of tests within the category. However, the success in diagnosing AD based on tests in this category was quite strong. Two of the combined scores (recall plus recognition, HVLT: Shi et al. 2012; CERAD: Sotaniemi et al. 2012), both demonstrated sensitivity and specificity scores into the 90%s. Another study in this category reported recognition span total (verbal, visual, facial), with excellent sensitivity and specificity values at 95% and 96%, respectively (Salmon et al. 1989). Unfortunately, this study did not report cutoff values. An interference score for the Fuld Object Memory Evaluation (Loewenstein et al. 2004) reported a specificity of 85% and sensitivity of 96%, again without a cutoff. As mentioned above, a factor score combining retention performance across multiple measures performed well, with 85% sensitivity and 92% specificity (Ivnik et al. 2000). Both prospective and retrospective components of a prospective memory test (Marcone et al. 2017) had high sensitivity (93 and 85%, respectively) and specificity (86 and 98%, respectively). Troster et al. (1993) reported sensitivity (88%) and specificity (99%) for the combined WMSR Logical Memory and Visual Reproduction percent savings scores, both well above the recommended 80% (Consensus Workgroup 1998). Finally, total recall of the Buschke Selective Reminding (combining recall from short-term and long-term memory) yielded a sensitivity of 95% and specificity of 100%, both well above the minimum suggested cutoff. Unfortunately, no cutoff values were reported (Paulsen et al. 1995). The results overall from this section suggest that using a combination of scores, particularly recall and recognition scores added together, could be beneficial in diagnosing AD. Future research should continue to examine combination scores with more uniformity. Normative data for such combined measures are needed.

## AD Versus Other Dementias/Disorders

Another important area of research focuses on the ability of neuropsychological measures to differentiate between AD and other neurological syndromes, including other dementias, neurological conditions and psychiatric disorders impacting neuropsychological functioning. The “other” category for this review was heterogeneous, and therefore unable to be included in meta-analysis. A total of 24 studies were initially identified, however, 14 were excluded according to exclusionary criteria described in the method. Online Resources (Table ii) presents the remaining 10 studies. These studies included comparisons between AD and semantic dementia, dementia due to Huntington’s disease (HD), Parkinson’s disease (PD), psychiatric populations, subcortical vascular dementia or small vessel disease (VaD), a sample of “non-AD” (described below), and Dementia with Lewy Bodies (DLB). In four of the nine studies, it was unclear if the test of interest was used to diagnose the disorder. Additionally, no studies in this section made 2 × 2 data available.

Four studies included measures of immediate memory and compared detection of AD to HD, VaD, semantic dementia, and “non-AD.” Three of these studies had acceptable sensitivity, including the RBANS Story Memory for AD versus VaD (McDermott and DeFilippis 2010), the Visual Route Learning Test for AD versus Semantic dementia (Pengas et al. 2010), and the Neuropsychological Assessment Battery (NAB) Daily Living Memory – Immediate recall (Gavett et al. 2012) for AD versus “non-AD.” Values for specificity were lower and more variable, but two studies (Gavett et al. 2012; Pengas et al. 2010) demonstrated both sensitivity and specificity values for immediate recall (NAB Daily Living Memory – Immediate Recall) that were above the recommended cutoff of 80%(Consensus Workgroup 1998). The Buschke Selective Reminding Test – Short-Term Memory (Paulsen et al. 1995) did not reach the recommended cutoff for either sensitivity or specificity in differentiating AD and HD.

Of nine delayed memory measures in six studies, five measures met the recommended cutoff for sensitivity, including Delayed Word Recall (O'Carroll et al. 1997) for AD versus Depression and the NAB Daily Memory Delayed Recall (Gavett et al. 2012) for AD versus “non-AD” group, as well as RBANS – List Recall (McDermott and DeFilippis 2010), RBANS – Delayed Memory Index (McDermott and DeFilippis 2010) and delayed figure recall for AD versus VaD (Matioli and Caramelli 2010). Only two of the nine tasks met the recommended cutoff for specificity, including an Enhanced Cued Recall task (Esen Saka and Elibol 2009) for AD versus PD, and the NAB Daily Living Memory – Delayed Recall (Gavett et al. 2012) for AD versus “non-AD.”

Three measures of recognition were included and all met the recommended cutoff of 80% (Consensus Workgroup 1998) for sensitivity, including Delayed Word Recognition (O'Carroll et al. 1997) for AD versus depression, RBANS-List Recognition (McDermott and DeFilippis 2010) for AD versus VaD, and CERAD or Auditory Verbal Learning Test (AVLT) List Recognition (Schmidtke and Hüll 2002) for AD versus small vessel disease. However, importantly, the specificity of recognition measures was low (ranging from 47 to 66%), for AD versus depression and VaD. A combined score of the WMS-R Logical Memory II and Object Assembly (Oda et al. 2009) for AD versus DLB had 81% sensitivity, and specificity falling just below the recommended cutoff at 76%. A combined measure reflecting recall from both short- and long-term memory on a selective reminding paradigm was below suggested cutoffs for differentiating dementia due to AD and HD (Paulsen et al. 1995). A second study (Troster et al. 1993) examined accuracy for differentiating mild AD from mild HD and moderate AD from moderate HD using the WMS-R Logical Memory plus Visual Reproduction percent savings. The moderate stage of the disease demonstrated good sensitivity (80%) and specificity (88%), whereas the mild stage of the disease had good sensitivity (86%) but poor specificity (36%).

In general for AD versus other comparisons studies found acceptable levels of sensitivity with low but varied specificity. Of note, the study with the highest reported specificity (Gavett et al. 2012) used a “non-AD” combined group that included healthy controls, MCI, dementia that was not AD, and ambiguous non-MCI. This enabled Gavett and colleagues to have a much larger sample relative to other studies and to evaluate specificity in terms of a broader neurological sample. Their findings suggest that although specificity for

Alzheimer's dementia may appear low when compared directly to other dementing conditions, relative to a broader neurological sample, differentiation based on objective memory scores fares well. Additionally, when considering individual measures, values for specificity or sensitivity may appear low, yet differential diagnosis in the clinical setting considers multiple measures and numerous factors in addition to neuropsychological test scores, likely resulting in better specificity for differential diagnosis than values based on a single memory score imply (see Fields et al. 2011 for discussion).

Importantly, clinicians often heavily weight recognition scores as a differential for AD compared to other dementias. Studies here indicate that it may be an erroneous assumption that non-AD populations have better recognition than AD patients. In addition to further exploration of the specificity of recognition or cued-recall paradigms, future research in the area of differentiating AD from other syndromes should also examine list learning immediate free recall, as no studies were found that included data for this type of measure.

## MCI Versus HC

The literature review and PubMed search yielded 60 studies that were deemed relevant based on our initial search criteria. After a more careful review of each study, 22 were excluded according to exclusionary criteria described in the method. Online Resources (Table iii) present the remaining 38 studies. Of the 38 studies, one study explicitly stated using the measure of interest in combination with other measures and diagnostic methods to diagnose MCI (Yassuda et al. 2010) and one study is presumed to have used the measure of interest to diagnose MCI (Karantzoulis et al. 2013). Additional studies have unclear diagnostic methods and do not explicitly state whether the measures of interest were used to diagnose individuals with MCI (Baek et al. 2012, 2011; Clark et al. 2010; Gavett et al. 2012; Karrasch et al. 2005; Lekeu et al. 2010; Lemos et al. 2014; Saka et al. 2006). Two studies by the same group of authors used separate samples of MCI participants (they report non-overlapping recruitment dates) but the same HC participants (Baek et al. 2012, 2011). Although we still report on these studies separately, the use of the same HC group may increase the similarity of diagnostic accuracy values between the studies. Most studies reported the cutoffs used to determine sensitivity and specificity values, however, seven studies did not (Lekeu et al. 2010; Loewenstein et al. 2004; Rabin et al. 2009; Serna et al. 2015; Shankle et al. 2005). Only two studies reported 2 × 2 data in addition to sensitivity and specificity values (Junkkila et al. 2012; Troyer et al. 2008). Almost all of the 38 studies (k = 36) used well-established MCI diagnostic criteria (Albert et al. 2011; Petersen et al. 1999; Petersen et al. 2001; Petersen 2007, 2004; Portet et al. 2006; Winblad et al. 2004). One study classified individuals with MCI based on NINCDS-ADRDA criteria for diagnosis of Alzheimer's disease, but without functional impairment (Loewenstein et al. 2006). A second study used the Clinical Dementia Rating (CDR) scale to diagnose MCI (Shankle et al. 2005).

Overall, when examining the data both qualitatively and quantitatively for a subset of the measures applied to the meta- analysis, sensitivity and specificity values for differentiating between individuals with MCI versus healthy elderly controls are lower than the suggested minimum cutoffs of 80% sensitivity and specificity recommended for differentiating AD patients from HC participants and other dementias (Consensus Workgroup 1998), and

generally lower than the values for AD versus HC reported in the present literature review and meta-analyses. We use the qualifier “adequate” to refer to sensitivity and specificity values ≥70%.

**Immediate Memory—**Seventeen data points contributed to the evaluation of immediate recall measures in differentiating MCI from HC (Table 4, Fig. 8). Overall, these measures demonstrated adequate diagnostic accuracy with values lower than the recommended minimum cutoff, but just above 70% for sensitivity (Se = .72, 95% CI [.63, .79]) and just at the cutoff recommended by the Consensus Workgroup (1998) for detecting AD for specificity (Sp = .81, 95% CI [.75, .85]) (Table 4). Visual inspection of the forest plots (Fig. 8) and SROC curves (Fig. 9) further suggests adequate diagnostic accuracy for differentiating MCI from healthy controls, but values are much lower than for AD vs HC comparisons.

To assess for potential bias in studies where the measures examined were also used to diagnose participants with MCI, we classified all 17 data points according to whether or not the measure was used to diagnose participants. A total of 13 studies did not use the measure under examination for diagnosis of participants, with 1 study using the measure for diagnosis and the remaining 3 studies being unclear as to how the measure was used in participant diagnosis. A meta-analysis of the 13 studies not using the measure for diagnosis was conducted (Table 5), indicating that the immediate memory recall measures continued to display adequate diagnostic accuracy for differentiating MCI from HC with values remaining lower than the suggested minimum cut-off for sensitivity (Se = .73, 95% CI [.63, .81]) and at the specificity cutoff recommended by the Consensus Workgroup (1998) for detecting AD (Sp = .80, 95% CI [.75, .85]). Visual inspection of the forest plots and SROC curves (Supplemental Figures xxi and xxii) supports this conclusion.

For immediate memory recall subclasses, all types of measures displayed at least adequate sensitivity and specificity (Table 4; List Free Recall Se = .72, 95% CI [.62, .81], Sp = .81, 95% CI [.75, .86]; Story Free Recall Se = .74, 95% CI [.50, .89], Sp = .74, 95% CI [.60, .84]; List Cued / Selective Reminding Se = .74, 95% CI [.54, .87], Sp = .84, 95% CI [.73, .90]). Visual inspection of the SROC curves for Immediate List Free Recall, Immediate Story Free Recall, and List Cued / Selective Reminding tests (Fig. 10) confirm that these measures display adequate (at least greater than 70%) diagnostic accuracy for differentiating MCI from healthy controls. Forest plots for subclass analyses are presented in Online Resource (Figures x–xii). Many of the studies reporting on immediate story recall also report sensitivity and specificity for immediate recall of word lists. In general, comparing story recall to list recall within each study revealed a trend of better sensitivity and specificity values for immediate list recall compared to immediate story recall (e.g., Baek et al. 2012; Blanco-Campal et al. 2009; Duff et al. 2010).

Three additional studies investigated measures of immediate memory that do not fall into any of the above mentioned categories. Loewenstein et al. (2006) investigated the diagnostic accuracy of the WMS-III immediate recall portion of Visual Reproduction and report 44% sensitivity and 91% specificity. Gavett et al. (2012) reports sensitivity (86%) and specificity (90%) values above the suggested minimal cutoff of 80% (1998) for the immediate recall

portion of NAB Daily Living Memory, which investigates memory for common daily information (e.g., medications, addresses). Duff et al. (2010) reported low sensitivity (35%) but good specificity (85%) on a combined memory score from the RBANS (immediate memory scores for list and story).

Overall, although the majority of studies that investigate the diagnostic accuracy of immediate memory measures do not exceed the suggested minimum cutoff of 80% put forth by the Consensus Workgroup (1998) for differentiating between AD and HC, many exceed 70% sensitivity and specificity.

**Delayed Memory—**Twenty-two data points contributed to the evaluation of delayed recall measures in differentiating MCI from HC (Table 6, Fig. 11). Overall, these measures demonstrated adequate diagnostic accuracy with values below, but approaching the recommended cutoff of 80% proposed by the Consensus Workgroup (1998) for sensitivity (Se = .75, 95% CI [.69, .81]) and just above the cutoff for specificity (Sp = .81, 95% CI [.77, .84]) (Table 6). Visual inspection of the forest plots (Fig. 11) and SROC curves (Fig. 12) further confirms adequate diagnostic accuracy of delayed recall measures for differentiating MCI from HC, with the values being lower than for AD vs HC comparisons. Comparing the diagnostic accuracy of delayed word-list recall to immediate wordlist recall within studies reporting both types of measures did not yield a consistent pattern with regards to one type of measure having higher diagnostic accuracy relative to another.

To assess for potential bias in studies where the measures examined were also used to diagnose participants with MCI, we classified all 22 papers according to whether or not the measure was used to diagnose participants. A total of 16 studies did not use the measure under examination for diagnosis of participants, with 1 study using the measure for diagnosis and the remaining 5 studies being unclear as to how the measure was used in participant diagnosis. A meta-analysis of the 16 studies not using the measure for diagnosis was conducted (Table 5), indicating that the delayed memory recall measures continued to display adequate diagnostic accuracy for differentiating MCI from HC with values remaining lower than the suggested minimum cut-off for sensitivity (Se = .76, 95% CI [.68, .82]) and at the cutoff recommended by the Consensus Workgroup (1998) for detecting AD for specificity (Sp = .81, 95% CI [.77, .85]). Visual inspection of the forest plots and SROC curves (Supplemental Figures xxiii and xxiv) supports this conclusion.

Forest plots for delayed memory recall subclasses are presented in Online Resources (Figures xiii–xvi, respectively). All measures display sensitivity values below the recommended cutoff (generally in the adequate range), although several specificity values fall above the recommended cutoff (List Free Recall Se = .73, 95% CI [.64, .81], Sp = .83, 95% CI [.77, .88]; Story Free Recall Se = .74, 95% CI [.56, .86], Sp = .79, 95% CI [.70, .85]; List Cued / Selective Reminding Se = .72, 95% CI [.58, .82], Sp = .85, 95% CI [.76, .91]; Visual Recall Se = .69, 95% CI [.33, .91], Sp = .82, 95% CI [.64, .92]). Visual inspection of the SROC curves for the delayed recall subclasses (Fig. 13) confirm that all measures display adequate diagnostic accuracy for differentiating MCI from HC. However, it is important to note that threshold effects were apparent for list cued / selective reminding (as evidenced by the forest plot and rho correlation). Comparing the diagnostic accuracy of

delayed story recall to immediate story recall in studies that examined both revealed an overall pattern of improved or comparable sensitivity and/or specificity for delayed story recall compared to immediate story recall.

There were an insufficient number of studies (<3) found for inclusion in a meta-analysis for a few types of measures. Of the two studies investigating the sensitivity and specificity of verbal retention scores (list or story), only one of the four measures (RBANS List Learning Retention; Clark et al. 2010) exceeded the suggested 80% cutoff (1998). The other tasks report specificity values above 80% but sensitivity values below 70%. Comparing measures within the same study revealed a pattern of superior sensitivity for recall scores relative to scores of retention belonging to the same measure (e.g. CERAD Word List; Blanco-Campal et al. 2009). Similarly, comparing within two studies (Lemos et al. 2015; Park et al. 2016) that investigated both immediate and delayed conditions of cued or selective reminding paradigms showed mixed results. Improved sensitivity and specificity values were reported by Park et al. (2016) for delayed conditions of the RI-24 (adapted from the RI-48) compared to immediate conditions. Lemos et al. (2015) reported comparable values between the immediate and delayed conditions of the FCSRT in Portuguese.

Within the “Other Delayed Recall” category, two of the three studies examined the Delayed Memory Index of the RBANS (Duff et al. 2010; Karantzoulis et al. 2013). Both report specificity that exceeds the 80% suggested minimum cutoff (Consensus Workgroup 1998), but sensitivity was variable, with only one (Karantzoulis et al. 2013) demonstrating adequate sensitivity (72%). Methodological differences between the two studies may account for this difference. Specifically, although not explicitly stated, it is presumed that Karantzoulis et al. (2013) used the RBANS to diagnose individuals with MCI. This circularity in methodology has the potential of inflating scores of sensitivity. Comparatively, Duff et al. (2010) explicitly did not use the RBANS to diagnose individuals with MCI and found much lower sensitivity on this measure (56%). Comparable with Gavett et al.’ (2012) examination of immediate recall of the NAB Daily Living Memory measure, the delayed recall trial yielded sensitivity and specificity values exceeding the suggested minimum cutoff of 80% (97% and 88%, respectively).

In summary, the values for sensitivity and specificity of delayed recall measures were significantly lower for differentiating MCI and HC than AD and HC, and often did not meet suggested minimum cutoffs (Consensus Workgroup 1998). However, a majority of the values reported within the meta-analyses as well as our qualitative review exceeded sensitivity and specificity levels of 70%.

**Associative Learning—**Eight studies reporting on fourteen different measures of associative learning were found through the literature search. These studies are qualitatively reviewed. Four measures exceeded suggested minimum cutoffs of the consensus report (1998). Wang et al. (2013) examined the Modified Spatial Context memory Test (SCMT).[8] The authors report strong sensitivity and specificity for the total score and the event-place association memory subtest (97% and 93% for total score, and 97% and 100% for subtest).

[8]This test examines associative memory of spatial location, event-place association, and place-object association memory.

A measure of associative memory investigated by Pike et al. (2013), the WMS-IV VPA delayed score, also exceeded suggested cutoffs for sensitivity and specificity. Finally, Troyer et al. (2008) report sensitivity and specificity for the Brief Visual Memory Test – Revised (BVMT-R) Object Location Recall[9] test (Benedict, 1997). In the case of the Troyer et al. study, an association score was derived separately from an accuracy score in order to examine associative learning independent of accuracy. The authors report 86% sensitivity and 97% specificity for the association score. Troyer et al. also examined sensitivity and specificity of Digit Symbol Incidental Recall.[10] They report specificity of 90% which exceeded the minimum cutoff of 80%, and sensitivity of 76%.

In summary, tasks of associative learning varied widely in their sensitivity and specificity for distinguishing between older adults with MCI and HC participants. Only eight of fourteen measures are reported to have sensitivity and specificity that both exceed even 70%. Overall, studies that investigated the sensitivity and specificity of the same measure in AD versus HC report higher values than MCI versus HC.

**Other Memory Measures—**This section is mainly comprised of studies that investigated the diagnostic accuracy of recognition and combined/interference scores. Additionally, three studies are included in the “miscellaneous” portion of the Online Resource (Table iii), two of which report on prospective memory measures (Blanco-Campal et al. 2009; Delprado et al. 2012) and one of which reports diagnostic accuracy scores stratified by education groups (Yassuda et al. 2010).

Six studies investigated the diagnostic accuracy of recognition measures. Only one study by Rabin et al. (2009) report sensitivity (92%) and specificity (84%) values that exceed the 80%suggestedminimum cutoff (Consensus Workgroup 1998), for the recognition portion of WMS-III Logical Memory, although the cut-off used was not reported. Only one of the remaining 5 studies report both sensitivity and specificity that exceed 70% (Fuld Object Memory Evaluation; Loewenstein et al. 2004). The remaining studies report sensitivity values of 74%or below and specificity values of 73%or below for wordlist recognition (Baek et al. 2012; Duff et al. 2010; Karrasch et al. 2005), story recognition (Baek et al. 2012), or photograph recognition (Ritter et al. 2006).

Overall, with the exception of two studies (Loewenstein et al. 2004; Rabin et al. 2009), recognition measures do not seem to provide strong diagnostic accuracy in distinguishing between MCI and HC groups. MCI versus HC findings varied with regards to sensitivity and specificity tradeoffs across recognition measures.

With regards to combined or interference scores, Shankle et al. (2005) investigated the diagnostic accuracy of a weighted score derived using correspondence analysis from the CERAD Word List test, reporting both sensitivity (94%) and specificity (89%) that exceed the minimum suggested cutoff. Of note, the authors did not report the cutoff score used to

[9]During this test, participants are asked to re-create a 2 by 3 array of six simple geometric figures after three 10 s learning trials. Points are awarded for accuracy and correct object placement.

[10]During this task participants are asked to recall the associated symbol of nine numbers that they initially learn through a coding task (WAIS-III Digit Symbol; (Wechsler 1997)).

derive these values, limiting the clinical applicability of the results. Two studies by Loewenstein et al. (2004, 2006) report sensitivity and specificity for the Fuld Combined Interference score. The first study by Loewenstein et al. (2004) reports sensitivity that nearly met the 80% minimum recommended cutoff (Consensus Workgroup 1998), while the second study (Loewenstein et al. 2006) reports lower sensitivity at 70%. Specificity was relatively comparable between the two studies (87% and 91% respectively). A study by Crocco et al. (2014) examined sensitivity and specificity of the Loewenstein-Acevedo Scales of Semantic Interference and Learning (LASSI-L) task, which involves free and cued recall of two different 15-item word lists (see Crocco et al. for full task description). The authors report high diagnostic accuracy for a combined score of List A and List B cued recall (88% sensitivity and 92% specificity).

Two studies included in the “Miscellaneous” category investigated the diagnostic accuracy of prospective memory measures. Blanco-Campal et al. (2009) report sensitivity and specificity values (84% and 95% respectively) that exceed the minimum suggested cutoff of 80% (Consensus Workgroup 1998) for one type of prospective memory score in which they asked participants to say a category any time they see a word that belongs to the category (Silly Sentences, Non-Specific condition). In a second condition in which participants are asked to say a category any time they see one specific word (Silly Sentences, Specific condition), the authors report lower sensitivity (74%) but equally high specificity (95%). Delprado et al. (2012) investigated two measures of prospective memory. Both scores on the Cambridge Prospective Memory Test (CAMPROMPT)[11] were found to have sensitivity and specificity ranging from 69 to 73%. A study by Yassuda et al. (2010) investigated differences in sensitivity/specificity of the Rivermead Behavioral Memory Test in Brazilian participants with greater than or less than 8 years of education. In general, scores are comparable across groups. Sensitivity is generally low to adequate but is only slightly lower in the <8 years education group than the >8 years education group (71% versus 69% respectively). Specificity is also relatively commensurate between the groups (profile: 81% and 79%, respectively).

## Discussion

Results revealed generally high sensitivity and specificity for AD versus HC comparisons with values that were well above the recommended 80% cutoff based on guidelines put forth by the 1998 consensus report of the Working Group on Molecular and Biochemical Markers of AD (1998). Reviewing measures that differentiated AD from other conditions yielded few studies and mixed results, with generally high sensitivity in the context of low or variable specificity. Examination of MCI versus HC studies revealed generally lower sensitivity and specificity across memory measures relative to that seen for AD versus HC comparisons.

### AD Versus HC

Meta-analytic results showed that measures of both immediate and delayed memory tasks consistently demonstrated high sensitivity and specificity values, especially those involving

[11]Involves three time-based and three event-based prospective memory items embedded within puzzles of attention (Delprado et al. 2012). Two scores can be derived for this measure, an event-based score and a time-based score.

verbal word list recall. It is possible that immediate memory may be sufficient to support objective evidence of memory impairment required for a clinical diagnosis of AD. Importantly, this is based on studies focusing only on the distinction between AD and HC. Studies included here typically have excluded other potential causes of memory impairment. Only a handful of studies reported sensitivity and specificity values for immediate and delayed conditions within the same study. Within those, there was some suggestion that story memory delayed recall may have higher overall diagnostic accuracy relative to immediate recall. Future studies directly comparing immediate versus delayed measures in the same sample of participants will help to elucidate the degree to which immediate memory measures may stand on their own as diagnostically useful tools in clinical evaluations, independent of delayed measures. It will be especially important to include non-AD groups in such studies to determine whether immediate memory can distinguish between alternative disease etiologies as well as delayed memory given past literature using null hypothesis testing that suggests delayed recall and recognition memory are important for differential diagnosis (Bondi et al. 1996; Delis et al. 2005; Tierney et al. 2001).

Meta-analytic findings of cued and selective reminding paradigms, visual free recall, delayed list retention and immediate story free recall also yielded sensitivity and specificity exceeding minimum suggested cutoff of 80% (Consensus Workgroup 1998). However, there were a small number of studies analyzed and caution is warranted in interpreting these data. Further research is necessary to expound upon these findings. Associative memory tasks also yielded promising findings that exceed the minimum suggested cutoff of 80% (Consensus Workgroup 1998). In contrast, qualitative examination of recognition memory tests frequently had low sensitivity. However, specificity values of most recognition memory tasks were >80%, thus in combination with more sensitive recall tasks, these may be clinically useful. Finally, combined scores (e.g., recognition + delayed recall) demonstrated excellent sensitivity and specificity in multiple studies (e.g. Shi et al. 2012; Sotaniemi et al. 2012).

### AD Versus Other Dementia/Disorders

Overall, studies investigating the ability of memory measures to differentiate between AD versus Other dementias/disorders yielded mixed results, with generally high sensitivity and variable specificity across studies. Important to note is that in making a differential diagnosis, clinicians consider numerous factors, including multiple neuropsychological measures, psychiatric measures, medical history, information from collateral sources, and imaging data. Thus, the specificity of a particular measure in conjunction with other sources is likely higher than the specificity of the measure when considered on its own (Fields et al. 2011).

Several other considerations arise when evaluating the combined results of these studies. First, the “other” category was heterogeneous across studies, covering conditions such as vascular dementia, Huntington’s disease, and psychiatric illnesses. Second, a variety of memory measures were reported across studies. Third, few studies overall were identified as belonging to this category in the present literature review. Given the limited number of studies, and variability between the populations investigated and the measures reported, we

were unable to perform meta-analyses. Furthermore, our ability to comment on, and generalize the results of the AD versus Other category based on qualitative inspection of the data is limited. Future research is needed to delineate the ability of measures to accurately distinguish between AD and other dementing conditions. Importantly, measures of diagnostic accuracy, as opposed to mean group comparisons alone, need to be included in these future studies.

### MCI Versus HC

Results of studies comparing MCI versus HC groups yielded a general pattern of lower sensitivity and specificity than values reported by studies differentiating between AD versus HC, and lower sensitivity values than the recommended cutoff of 80% put forth by the Consensus Workgroup (1998) that are specifically suggested for AD versus HC comparisons. Of note, meta-analytic results yielded sensitivity and specificity values for all classes of memory measures ≥70%, and qualitative review of other memory subtypes similarly generally exceeded this level.

There are several potential explanations for the lowered sensitivity and specificity of neuropsychological memory measures in distinguishing between MCI and HC compared to AD versus HC. Individuals with MCI are less cognitively impaired than individuals with AD, thus potentially lowering the ability of measures to accurately distinguish between healthy and MCI groups. This may, in part, reflect the psychometric properties of some tests, with some displaying ceiling effects when used in cognitively intact samples. This limitation of test construction may contribute to a reduced sensitivity and specificity for detection of more mild or subtle forms of memory impairment. Thus, developing tests with heightened sensitivity and specificity for subclinical impairments is of critical importance.

The nature of the MCI construct and the variability with which MCI is conceptualized and diagnosed across studies likely also contributes to the lower diagnostic accuracy values in the MCI versus HC comparison. Complicating matters is that studies reported in the MCI versus HC comparison varied with regards to their level of detail in explaining their diagnostic methods and specific sample characteristics. Although studies reported here generally used well-established MCI diagnostic criteria, implementation of these criteria may differ widely between studies even when using the same diagnostic criteria. For example, while some studies provided specific information regarding the cutoff of impairment used to classify someone as having objective cognitive impairment (e.g., at least one memory measure <1.5 SD), others only generally stated that they followed the accepted diagnostic criteria (e.g., Petersen et al. 1999) without providing more detail. Similarly, some studies provided details regarding their methods for assessing functional abilities (e.g., activities of daily living assessments) and subjective complaints, while others did not. This variability between studies lends to the challenge in making cross-study comparisons of measures. Future studies should explicitly report the methods for diagnosing groups. This includes discussing the assessment measures used in the diagnosis and reporting the cutoff of impairment used to classify someone as impaired on a measure or cognitive domain.

Related to this, variations in the cutoffs and required numbers of impaired test scores used to diagnose MCI have resulted in large differences in the prevalence of the disorder (Ganguli et

al. 2011; Jak et al. 2009). In addition, some individuals diagnosed with MCI revert back to normal, or never convert to AD (Klekociuk et al. 2014; Mitchell and Shiri-Feshki 2009; Summers and Saunders 2012), thus suggesting that these individuals may represent false positives and are not representative of early Alzheimer's dementia (Edmonds et al., 2015a). A potential solution emerging from recent research is to use more lenient cutoffs (e.g., <−1 SD), but to do so across multiple cognitive measures (Jak et al. 2009; Klekociuk et al. 2014; Summers and Saunders 2012). A recent meta-analysis by Callahan and colleagues supports this conclusion. They found that impairment (defined as <−1SD) on two episodic memory measures predicted AD with 75.91% accuracy and the authors recommend using these cutoffs in classifying episodic memory impairments in MCI. In fact, the requirement that multiple tests must evidence impairment (albeit at a lower cutoff of <−1 SD) has resulted in more stable and reliable diagnoses (Bondi et al. 2014; Clark et al. 2013) because this decisional strategy obviates the base rate problem in aging populations (e.g., the rate of impaired test scores in neurologically normal populations; see Binder et al. 2009; Brooks et al. 2007).

Another source of variability lies in the etiological heterogeneity of MCI diagnoses. Individuals with MCI represent a heterogeneous group with multiple underlying etiologies accounting for cognitive impairment (Albert et al. 2011; Dubois et al. 2010; Dubois et al. 2007). In studies that we reviewed, 28 studies explicitly stated the subtype of MCI under investigation (e.g., amnestic MCI, multiple subtypes). Of these 28 studies, four included participants with non-amnestic MCI. The remaining ten studies were more general and it is unclear whether they limited their findings to amnestic MCI (aMCI) or all MCI subtypes. This is an important distinction given the increased likelihood of aMCIs developing AD compared to non-amnestic subtypes (Fischer et al. 2007), and the lower likelihood that non-amnestic MCI individuals will differ from controls on memory measures. Collapsing across MCI subtypes may result in a sample of individuals with varying etiologies, some of whom may develop other dementing conditions. This inevitably dilutes the diagnostic accuracy of detecting MCI using memory measures (Petersen 2004). Reporting the subtypes included in the MCI sample, or ideally stratifying results by subtype, will allow for more accurate representation of a measure's diagnostic utility. Additionally, biomarker support of an Alzheimer's etiology or longitudinal designs that confirm future cognitive decline or eventual progression to AD will provide more accurate diagnostic accuracy statistics. In general, the studies presented here did not provide biomarker confirmation of MCI due to AD. Importantly, only four (De Jager et al. 2003; Lekeu et al. 2010; Rabin et al. 2009) of the 38 studies reviewed provided follow-up information regarding rates of progression from MCI to AD. None of these four studies report differences in diagnostic accuracy of measures stratified by "converters" and "non-converters," although one study reports sensitivity and specificity values for predicting progression to AD (Rabin et al. 2009). Future studies with longitudinal designs and adequate power should do this as it would strengthen the validity of the diagnostic accuracy findings.

## Qualitative Patterns across Types of Memory Measures

Although it is problematic to make relativistic comparisons across memory measures based on meta-analytic results given that different samples, methodologies, and measures were

used across the studies, we highlight qualitative patterns we observed across measures and diagnostic categories in order to summarize the vast number of studies reported in the present review. Such comparisons should be interpreted with caution as they may be overly simplistic, although our qualitative review results above include discussion of these comparisons within the same study when possible. Within the AD versus HC studies, measures of immediate and delayed recall showed relatively equal sensitivity and specificity values (>80%). Comparisons of measures of immediate and delayed recall within MCI versus HC studies that examined both types of measures revealed a subtle trend for comparable or higher sensitivity and/or specificity for delayed recall measures compared to immediate recall. Word list recall scores generally showed higher diagnostic accuracy relative to story recall, retention scores, and recognition measures in both MCI and AD when compared within the same study (Baek et al. 2012; Blanco-Campal et al. 2009; Duff et al. 2008; Salmon et al. 2002).

Cued and selective reminding paradigms also had above-cutoff (i.e., ≥80%) sensitivity and specificity for differentiating between AD and HC. Although fewer studies examined these tasks in MCI and HC comparisons, studies that did generally showed adequate sensitivity and/or specificity (i.e., ≥70%). Carlesimo et al. (2011) also cited encouraging results in support of the diagnostic utility of cued and selective reminding paradigms in differentiating between AD and other forms of dementia, as well as their ability to predict MCI to AD conversion. In contrast, in a broad review of measures that predict conversion from MCI to AD, Gainotti et al. (2014) discuss inconsistent findings in the literature regarding the ability of cued recall tests to discriminate between MCI and AD patients, and MCI and HC participants. The authors conclude that cued recall paradigms are not necessarily better at predicting conversion from MCI to AD than measures of free recall. One possible reason for differences between studies is that cued paradigms are often vulnerable to ceiling effects, thus increasing the likelihood that those performing even slightly below ceiling renders an “impaired” performance and in turn improving sensitivity and specificity of these measures in some instances (AD versus HC) but not others (MCI versus HC). Given promising findings from the present review, more studies are certainly needed to evaluate the diagnostic accuracy and predictive utility of cued recall and selective reminding paradigms.

Associative learning tasks also emerged as having above-cutoff (≥80%) sensitivity and specificity in differentiating between AD and HC. This was generally not the case for studies investigating associative learning in MCI versus HC, although there is substantial heterogeneity of the measures reported in this subcategory and thus directly comparing across studies of associative memory is inherently problematic. Nevertheless, one possible conclusion is that associative learning may be less clinically useful for detecting early cognitive changes associated with AD. However, as discussed above, it is possible that the MCI studies presented also included individuals with non-Alzheimer’s etiologies (e.g., individuals with non-amnestic MCI who by definition do not have a memory impairment), thus making it more difficult to differentiate between MCI and HC with memory tests. Currently, there are studies (Troyer et al. 2012; Turriziani et al. 2004; Yonelinas et al. 2001) pointing to the relationship between associative learning and known early changes in medial temporal lobe functioning and related structures. Based on this strong theoretical base and current findings showing some measures of associative memory with high sensitivity and

specificity for differentiating MCI and HC, more research into associative learning as a helpful tool for early detection of preclinical AD is needed (see Lowndes and Savage 2007 for review). The diagnostic accuracy of recognition memory in MCI versus HC also fell below recommended cutoffs. Either more sophisticated and difficult recognition paradigms need to be implemented within this population, or results should focus on measures of recall for pre-dementia groups.

## Role of Neuropsychology as a Cognitive Biomarker of AD

The current results highlight the utility of neuropsychological measures of memory as valid cognitive biomarkers of AD. Neuropsychological tests appear to meet all of the features of an ideal diagnostic test, as put forth by the 1998 Consensus Workgroup. Specifically, they are largely reliable (but see Calamia et al. 2013 for exceptions), non-invasive, simple to perform, inexpensive, and neuropathologically validated (e.g., Kanne et al. 1998; Naslund et al. 2000). Furthermore, the results of this review confirm that many neuropsychological tests of memory are also precise. They are highly sensitive and specific in differentiating between patients with AD and healthy elderly controls. The ranges of sensitivity and specificity reported by many of the studies reviewed are comparable if not superior to those of other validated biomarkers. A review and meta-analysis by Bloudek et al. (2011) reported sensitivity and specificity values for widely accepted biomarkers. Studies reviewed included cerebrospinal fluid (CSF) biomarkers such as $A\beta_{42}$, phosphorylated tau, total elevated tau, $A\beta_{42}$ + Tau, and imaging methods included Fluorodeoxyglucose (FDG) Positron Emission Tomography (PET), single photon emission computed tomography (SPECT), and magnetic resonance imaging (MRI) studies. The authors report sensitivity values ranging from 80% to 90% and specificity values ranging from 82% to 90% in differentiating AD from non-demented controls. These values are in line with many of the memory measures reported in the current review for differentiating between AD and HC, especially immediate and delayed list free recall. With regards to differentiating AD from non-AD or other dementias, Bloudek et al. (2011) report sensitivity ranging from 73% to 93% and specificity ranging from 67 to 81%. The specificity of these biomarkers may be higher for differentiating AD from other dementias relative to memory measures based on the limited studies available in this review. Similar to what has been outlined in the appropriate use criteria for amyloid imaging (Johnson et al. 2013a, b), this finding suggests a potential clinical role for biomarkers when clinical history and neuropsychological testing are atypical for a suspected AD process. Importantly, these appropriate use criteria highlight that the clinical use of biomarker studies such as amyloid imaging is inappropriate if the cognitive complaint has not been objectively confirmed, ideally through neuropsychological evaluation.

Diagnostic imaging techniques and CSF biomarkers are not without their limitations. CSF biomarkers and PET imaging are invasive and expensive. McKhann et al. (2011) discuss their reasoning for not advocating the use of AD biomarker tests for routine diagnostic purposes, mainly citing the limited standardization of biomarkers across clinical settings and limited access to biomarkers across community settings. Additionally, one of the most commonly used tracers in research studies, (11)carbon-labeled Pittsburgh Compound-B ($^{11}$C–PIB-PET) that is thought to bind to brain fibrillar Aβ deposits, has a short half-life, limiting its clinical use. As a result, several alternative radiopharmaceuticals have been

developed, including [F-18] florbetapir. Florbetapir, which was approved by the Food and Drug Administration (FDA) in April 2012 and has been shown to correlate highly with $^{11}$C–PIB-PET (Johnson et al. 2013a, b).

Cerebro-spinal fluid and imaging biomarkers are also subject to false positives (low specificity). This is especially true for amyloid imaging, in which a positive scan alone does not indicate definite early Alzheimer's disease. One review study (Perani et al. 2013) found Amyloid PET to have an average sensitivity of 91%, with specificity values ranging widely (41–99%), suggesting that a positive PET scan does not necessarily indicate AD (high false positives). Consistent with this notion, Petersen et al. (2013) reported that none of the MCI subjects with only a positive amyloid scan (without evidence of neurodegeneration) progressed to dementia after 12–15 months. A recent case study reported a positive amyloid scan in a patient with dementia but without any neurofibrillary tangles or amyloid plaques at autopsy (Ducharme et al. 2013), highlighting that other etiologies, particularly cerebral amyloid angiopathy and Lewy bodies (Edison et al. 2008), can underlie a positive scan. Also, a recent Cochrane systematic review (Zhang et al. 2014) that used a reference standard of progression from MCI to Alzheimer's dementia concluded that $^{11}$C–PIB-PET, despite showing promise, could not be recommended for routine use in clinical practice given methodological variation across studies.

Neuropsychological testing has several benefits over CSF biomarkers and diagnostic imaging. Aside from testing fatigue, neuropsychological testing is non-invasive and has minimal negative effects. Additionally, neuropsychological testing is more accessible and less expensive than many of the CSF and imaging biomarkers currently available to clinical settings. Another advantage of neuropsychological testing is that a negative neuropsychological exam can rule out MCI or AD, and a positive exam in combination with a typical clinical neurological exam is often sufficient to diagnose AD and MCI (i.e., the core clinical criteria described by McKhann et al. 2011 and Albert et al. 2011, respectively). Supplemental biomarkers in more difficult or unusual cases may be helpful, for example if there are confounds (e.g., low education, cultural factors) or the possibility for a non-AD dementia (Johnson et al. 2013a, b; Laforce and Rabinovici 2011). In contrast, individuals with normal neuropsychological profiles can be found to be amyloid positive based on amyloid PET or CSF studies and many individuals with pathological diagnoses of AD at autopsy do not carry a clinical diagnosis in life (Davis et al. 1999; Hardy and Selkoe 2002). In addition, debate continues as to whether amyloid alone is sufficient for the development of AD (Herrup 2015). Positive amyloid findings (or even the general disclosure of one's genetic susceptibility via APOE genotype) can be very distressing for individuals (Lineweaver et al. 2014), and may cause undue harm and iatrogenic effects. For ethical reasons and psychological well-being, until there are available treatments for individuals who are amyloid positive that can prevent further progression, biomarker testing is suggested to be reserved for individuals with neuropsychological evidence of MCI or dementia.

Like CSF or plasma biomarkers, neuropsychological tests have also been shown to be sensitive to differences in underlying pathologies between dementing conditions. For example, patterns of performance across neuropsychological tests, as well as within the

same test (e.g., CVLT, Delis et al. 1991), have been shown to differ in patients with cortical dementias such as AD and subcortical dementias such as vascular dementia (e.g., Delis et al. 1991; Graham et al. 2004; Looi and Sachdev 1999). However, these studies examined differences in mean performance between groups and did not report diagnostic accuracy statistics. In our review of AD versus Other studies, many memory measures were adequate at identifying AD (high sensitivity) but weaker at differentiating AD from non-AD cases (low specificity), suggesting that memory measures may not provide the strongest diagnostic utility when differentiating between dementia types. Again, it is important to note that few studies were reviewed in the AD versus Other category and more research is needed to make generalizations regarding the ability of different neuropsychological measures of memory and other cognitive domains to accurately distinguish between AD and other dementias.

As the knowledge of useful biomarkers in AD rapidly expands, clinical trials for individuals with MCI or preclinical AD who are biomarker positive are already underway. Inclusion of these individuals helps to increase the confidence that participants included in the clinical trial will eventually progress to AD. Doing so is critical as clinical trials seek to alter the pathophysiological trajectory of Alzheimer's disease. In the studies reviewed presently, none of the MCI versus HC studies and only one of the AD versus HC studies (Ewers et al. 2012) included CSF or imaging biomarkers as part of their diagnostic accuracy reports. Nevertheless, neuropsychological measures may also play a role in preselecting individuals at highest risk for future development of Alzheimer's dementia. Currently, the FDA support the use of cognition as an outcome measure in clinical trials of preclinical AD (Food and Administration 2013). Identification of early cognitive markers that can assist in targeted recruitment of these individuals and use of reliable cognitive outcomes to precisely track their cognitive course is paramount. Results of the present review may be helpful for identifying the types of memory measures that may hold promise for detection of preclinical AD and for tracking cognitive change over time, and for avoiding measures with low sensitivity to MCI and AD that would likely perform poorly for this purpose. Future studies may consider combining CSF or imaging biomarkers with neuropsychological memory measures, which may further strengthen the confidence that individuals selected for clinical trials are at high risk of developing AD.

## Recommendations

In order to advance the field of clinical neuropsychology and highlight the utility of neuropsychological measures as cognitive biomarkers of AD, future neuropsychological studies should emphasize diagnostic validity statistics in addition to, or in place of, null hypothesis testing. This will promote the ongoing use of routine neuropsychological testing to aid in early identification and diagnosis of MCI and AD, as the broader Alzheimer's research field incorporates and relies upon biomarkers with increased frequency. Over the past 15– 20 years, there has been a marked increase in the number of studies of neuropsychological measures that incorporate diagnostic accuracy statistics into their findings. As this continues to become increasingly commonplace, it is important that studies follow important standards for reporting diagnostic accuracy statistics. Through the STARdem initiative (Standards for the Reporting of Diagnostic Accuracy Studies, international consensus process on reporting standards in dementia and cognitive

impairment), Noel-Storr et al. (2014) discuss guidelines, encompassed within four central areas, to which studies reporting diagnostic statistics should adhere. With regards to the first area, “study population,” the authors suggest that reports should address whether their sample was representative of the larger population, otherwise the test accuracy may be over- or underestimated. The second area refers to the “reference standard.” The authors discuss that inconsistently applied reference standards can lead to difficulty effectively evaluating performance of a test across studies. Third, “circularity” refers to use of the test under evaluation as part of the reference standard. At the very least, the authors recommend that research reports should make clear that such bias exists in their study. Finally, test-retest “reliability” of measures should be made clear in research reports of diagnostic accuracy statistics. The authors provide a 25-item checklist that further breaks down these four areas (see Noel-Storr et al. 2014).

Many of the 76 studies reported in the current review failed to meet one or more of the guidelines recommended by STARDem (Noel-Storr et al. 2014). Approximately 7% of the total studies reviewed (5/76 studies) used the measure of interest as part of the reference standard, introducing circularity and potentially inflating sensitivity and specificity findings. An additional 21% (16/76 studies) did not specify whether or not they did so, thus making it difficult to know whether values reported are truly representative. However, primary meta-analytic results did not change when we restricted the current meta-analysis to those studies that specifically did not use the measure of interest for diagnosis, which suggests that incorporation bias did not drive the current findings. Regardless of whether the specific index memory measure was also used as a reference measure for diagnostic purposes, the issue of circularity is still relevant for studies investigating diagnostic accuracy of memory measures. Future studies would also benefit from additional investigation of the potential influence of study quality on results of similar qualitative reviews and meta-analyses by incorporating formal checklists such as the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2; Whiting et al. 2011) to better quantify study quality (Reitsma et al. 2009).

Related to STARDem’s recommendation for research reports to include test-retest reliability information, some memory measures have less than optimal test-retest reliability (see Calamia et al. 2013). This is an important consideration, as lower reliability can place a ceiling on validity and reduce statistical power, thus attenuating classification accuracy statistics (Dennis et al. 1997; Ellis 2010). Consequently, it is possible that the diagnostic accuracy values reported in this study are deflated due to suboptimal reliability of some measures. At the very least, studies reporting classification accuracy should discuss the test-retest reliability of their measures of interest. This was not done for most of the studies reported in the present review. More broadly, neuropsychologists should consider the importance of test-retest reliability when developing new assessment tools. Improving reliability of memory measures may yield improved classification accuracy and validity coefficients.

Also important in studies of diagnostic accuracy is justification and reporting of the cut-off used (Noel-Storr et al. 2014), either optimal or conventional, to derive the sensitivity and specificity values. In our review of the literature, 14/76 (18%) of the studies across the three

comparisons did not report cut-off scores. This limits the clinical applicability and replication of such findings. Relatedly, many studies reported optimal cut-offs obtained through ROC analyses, allowing for optimization of sensitivity and specificity values. Although informative in many ways, this limits the generalizability of findings and their clinical applicability. Studies that implemented conventional cutoffs have the benefit of being clinically applicable and replicable in future research reports. Also potentially limiting generalizability and clinical applicability is the fact that sensitivity and specificity values do not reflect base rates of disorders under consideration. Sensitivity and specificity values should only be applied to diagnostic classification decisions at the individual level in the context of relevant base rates of the condition of interest by converting sensitivity and specificity values to positive and negative predictive values (see Gavett et al. 2012; Smith et al. 2008 for review).

Another consideration for future studies reporting on diagnostic accuracy of neuropsychological measures is including sensitivity and specificity as a key term. As part of their recommended checklist for reporting of diagnostic accuracy in studies of dementia, STARDem (Noel-Storr et al. 2014) recommend utilizing the medical subject heading (MeSH) of “sensitivity and specificity.” This ensures that articles are indexed appropriately and located with ease. A challenge of the present review was locating all appropriate articles through our PubMed search. It is possible that the search terms utilized in the PubMed/MEDLINE search missed relevant articles. In addition, we only used one database to search for articles, and several studies ultimately identified through other means were not returned in our database search. To minimize these limitations, we additionally conducted an independent record review based on prior knowledge, additional search terms and scanning of reference lists in identified studies. Nevertheless, the possibility remains that due to our selected search terms, studies were overlooked. For example, incorporating a term such as *memory* in place of Neuropsychology may yield additional studies that did not specify *neuropsychology* as a key word. The decision to use *neuropsychology* as a search term was made in an effort to reduce the number of studies using only screening measures, as such measures were not a target of the current review.

## Other Study Limitations and Future Directions

The vast majority of studies in this review report cross-sectional data and only a minority of the studies reviewed included participants who were followed longitudinally. For the purposes of the current review, only baseline data of these longitudinal studies were examined, minimizing the effect of practice on our data. Although we did not report on longitudinal studies that investigate the predictive utility of neuropsychological measures, diagnostic accuracy studies of MCI and AD would be strengthened by including longitudinal data regarding conversion rates of MCI to AD and further confirmation of AD etiology through pathological or biomarker confirmation. This would strengthen the confidence that sensitivity and specificity values accurately reflect true diagnosis of underlying Alzheimer’s disease. As mentioned above, only three of the MCI studies reported include information on progression. Surprisingly, only three studies included in the current review included pathological or genetic confirmation of AD (Parra et al. 2010; Salmon et al. 2002; Storandt and Morris 2010). Increased etiologic certainty in studies would likely further increase the

diagnostic accuracy values of memory measures. On the other hand, the notion of “pure” AD pathology is increasingly suspected to be more rare than multiple underlying neuropathologies (e.g., AD, cerebrovascular disease, Lewy bodies, hippocampal sclerosis, TDP-43) for the vast majority of late-onset and sporadic forms of AD (Schneider et al. 2009; Zlokovic 2011).

Studies with longitudinal data can also investigate the ability of neuropsychological tests to predict conversion from preclinical AD to MCI and from MCI to AD. Research has consistently shown that many individuals with MCI remain stable over time, never progressing to AD, while some may even revert back to “normal” (Klekociuk et al. 2014; Mitchell and Shiri-Feshki 2009; Summers and Saunders 2012; Winblad et al. 2004). As discussed above, part of the challenge in conducting research with MCI as a diagnostic category is the fact that MCI is a heterogeneous construct, with distinct subtypes of MCI potentially representing variations in the underlying etiology (Petersen et al. 2009; Winblad et al. 2004). Additionally, although the vast majority of research points to specific subtypes of MCI (e.g., single- and multidomain amnestic MCI) as being more likely to have underlying AD (e.g., Han et al. 2012;Ravaglia et al. 2005; Yaffe et al. 2006), other research has revealed that non-amnestic subtypes also develop AD (e.g., Busse et al. 2006; Fischer et al. 2007). Again, the notion of ‘pure’ AD, rather than the common possibility of multiple underlying neuropathologies, furthers this ambiguity. Thus, identifying measures that can accurately predict progression will prove valuable for both research and clinical purposes, in order to identify those individuals across all MCI subtypes and neuropathologic substrates.

Often, studies that investigate the predictive utility of neuropsychological measures apply null hypothesis testing rather than report on the diagnostic accuracy of measures. Furthermore, many of the predictive studies that have reported diagnostic accuracy statistics utilize composite measures or multiple measures in one predictive model (Gainotti et al. 2014), thus making it difficult to judge the utility of any single neuropsychological measure. Although we did not include prediction studies in the present review, a recent review by Gainotti et al. (2014) found that measures of delayed recall are the best neuropsychological predictors of conversion from MCI to AD, with sensitivity values ranging from 73 to 75% to 86–89% and specificity ranging from 70% to 94–97% in studies reviewed. Our review of studies differentiating between MCI and HC also found that delayed recall measures outperformed other types of memory measures with regards to sensitivity and specificity, suggesting that these measures of episodic memory are particularly sensitive to early AD changes.

Many recent longitudinal studies have directly compared the ability of cognitive measures versus biomarkers at baseline to predict progression over time, or evaluated the utility of multiple markers for predicting progression. Across studies, there is growing support that when examining individual predictors, neuropsychological measures are the best predictors of progression and conversion to Alzheimer’s dementia (Eckerström et al. 2013; Gomar et al. 2014), or at the very least perform equally as well as various biomarkers (Ewers et al. 2012; Palmqvist et al. 2012). In addition, a combination of neuropsychological measures and biomarkers typically outperforms any individual predictor (Devanand et al. 2008; Eckerström et al. 2013; Heister et al. 2011; Landau et al. 2010; Peters et al. 2014).

Unfortunately, because multiple measures are typically investigated simultaneously in these studies, it is difficult to ascertain the sensitivity and specificity for predicting conversion of each individual measure.

The present review was limited to neuropsychological measures of episodic memory. Other cognitive domains may also prove useful in differentiating between individuals at risk for or diagnosed with AD and healthy elderly controls or other dementias. Measures of semantic memory have been shown to detect early AD changes (Gainotti et al. 2014) and may be useful in differentiating MCI and AD from other diagnostic groups and from healthy elderly controls. Additionally, there is growing evidence pointing to measures of attention, processing speed, and executive functioning as important for preclinical detection of AD (see Rentz et al. 2013 for review). These measures may be particularly useful in detecting non-amnestic subtypes of MCI and their progression to AD.

The present review did not assess for publication bias, as methods for computing publication bias for studies of diagnostic accuracy have questionable or unknown validity (Macaskill et al. 2009). The current review also did not evaluate the potential role of several important covariates that can impact diagnostic accuracy and may have varied across studies, including age, sex, education, race and ethnicity. Future studies should consider examining the effect of these variables on diagnostic accuracy. In addition, we included studies that used either raw scores or demographically-corrected standardized scores for their cut-offs, and Tables i–iii indicate which type of score was used for each study. Further, studies used both optimally-derived cut-offs that are more study specific and tended to represent raw score cutoffs and conventional cut-offs that were more frequently based on normative scores. However, we did not examine for any potential differences in diagnostic accuracy statistics by the type of score or cut-off used, which may be an important future direction given evidence that raw scores may be more sensitive and demographically-corrected scores more specific (O'Connell and Tuokko 2010). Future studies are needed to compare the diagnostic accuracy of these different types of scores and cutoffs to determine which yield the highest levels of sensitivity and specificity when compared within the same study.

## Concluding Comments

With an increasing focus on biomarkers in the field of dementia research, it is important to highlight the role of neuropsychological testing in detecting AD. Findings reveal that many measures of memory meet the suggested sensitivity and specificity guidelines put forth by the Consensus Workgroup (1998) for biomarkers in differentiating between patients with AD and HC. Diagnostic accuracy values for differentiating between MCI and HC are also promising but require further refinement. Future research focusing on specific MCI subtypes, including biomarkers to support a diagnosis of MCI due to Alzheimer’s disease, implementing longitudinal datasets, and investigating the predictive utility of neuropsychological measures will help strengthen our understanding of the role of neuropsychological measures as ideal diagnostic tests in preclinical AD, MCI, and Alzheimer’s dementia.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research was partially supported by the Department of Veterans Affairs, Veterans Health Administration, VISN 1 Career Development Award (N.H.S.), as well as National Institutes of Health grants R01 AG049810 (M.W.B.) and K24 AG026431 (M.W.B.). This material is also based upon work supported by the Office of Academic Affiliations, Department of Veterans Affairs. NHS serves as a consultant to Biogen. MWB serves as a consultant to Novartis and Eisai and receives royalties from Oxford University Press. KS was supported in this work by a University of the Sunshine Coast Research Scholarship. MS reports personal fees from Eli Lily (Australia) Pty Ltd and grants from Novotech Pty Ltd, outside the submitted work. The contents of this article do not represent the views of the U.S. Department of Veterans Affairs of the United States Government.

## References

*. Adam S, Van der Linden M, Ivanoiu A, Juillerat AC, Bechet S, Salmon E. Optimization of encoding specificity for the diagnosis of early AD: The RI-48 task. Journal of Clinical and Experimental Neuropsychology. 2007; 29(5):477–487. [PubMed: 17564913]

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment 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):270–279.

Alegret M, Rodriguez O, Espinosa A, Ortega G, Sanabria A, Valero S, et al. Concordance between subjective and objective memory impairment in volunteer subjects. Journal of Alzheimer's Disease. 2015; 48(4):1109–1117.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition - Text Revision (DSMIV-TR) (IV ed.). Washington D.C: American Psychiatric Publishing, Inc; 2000.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub; 2013.

American Academy of Neurology. Assessment: Neuropsychological testing of adults. Considerations for neurologists. Neurology. 1996; 47:592–599. [PubMed: 8757049]

*. Baek MJ, Kim HJ, Ryu HJ, Lee SH, Han SH, Na HR, et al. The usefulness of the story recall test in patients with mild cognitive impairment and Alzheimer's disease. Aging, Neuropsychology, and Cognition. 2011; 18(2):214–229.

Baek MJ, Kim HJ, Kim S. Comparison between the story recall test and the word-list learning test in Korean patients with mild cognitive impairment and early stage of Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology. 2012; 34(4):396–404. [PubMed: 22263656]

Beglinger LJ, Duff K, Moser DJ, Cross SA, Kareken DA. The Indiana faces in places test: Preliminary findings on a new Visuospatial memory test in patients with mild cognitive impairment. Archives of Clinical Neuropsychology. 2009; 24(6):607. [PubMed: 19679593]

Benedict, RHB. Brief Visuospatial memory test-revised. Lutz: Psychological Assessment Resources, Inc; 1997.

*. Bertolucci PHF, Okamoto IH, Brucki SMD, Siviero MO, Toniolo Neto J, Ramos LR. Applicability of the CERAD neuropsychological battery to Brazilian elderly. Arquivos de Neuro-Psiquiatria. 2001; 59(3A):532–536. [PubMed: 11588630]

Binder LM, Iverson GL, Brooks BL. To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. Archives of Clinical Neuropsychology. 2009; 24:31–46. [PubMed: 19395355]

*. Blanco-Campal A, Coen RF, Lawlor BA, Walsh JB, Burke TE. Detection of prospective memory deficits in mild cognitive impairment of suspected Alzheimer's disease etiology using a novel event-based prospective memory task. Journal of the International Neuropsychological Society. 2009; 15(1):154–159. [PubMed: 19128540]

Bloudek LM, Spackman DE, Blankenburg M, Sullivan SD. Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. Journal of Alzheimer's Disease. 2011; 26(4):627–645.

Bondi MW, Smith GE. Mild cognitive impairment: A concept and diagnostic entity in need of input from neuropsychology. Journal of the International Neuropsychological Society. 2014; 20(02): 129–134. [PubMed: 24490866]

Bondi, MW., Salmon, DP., Kaszniak, A. The neuropsychology of dementia. In: Grant, I., Adams, KM., editors. Neuropsychological assessment of neuropsychiatric disorders. 2. New York: Oxford; 1996. p. 164-199.

Bondi MW, Edmonds EC, Jak AJ, Clark LR, Delano-Wood L, McDonald CR, et al. Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. Journal of Alzheimer's Disease. 2014; 42(1):275–289.

Brooks BL, Iverson GL, White T. Substantial risk of “accidental MCI” in healthy older adults: Base rates of low memory scores in neuropsychological assessment. Journal of the International Neuropsychological Society. 2007; 13(03):490–500. [PubMed: 17445298]

Busse A, Hensel A, Guhne U, Angermeyer MC, Riedel-Heller SG. Mild cognitive impairment: Long-term course of four clinical subtypes. Neurology. 2006; 67(12):2176–2185. [PubMed: 17190940]

*. Cahn DA, Salmon DP, Butters N, Wiederholt WC, Corey-Bloom J, Edelstein SL, et al. Detection of dementia of the Alzheimer type in a population-based sample: Neuropsychological test performance. Journal of the International Neuropsychological Society. 1995; 1(3):252–260. [PubMed: 9375219]

Calamia M, Markon K, Tranel D. The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. The Clinical Neuropsychologist. 2013; 27(7):1077–1105. [PubMed: 24016131]

Carlesimo GA, Perri R, Caltagirone C. Category cued recall following controlled encoding as a neuropsychological tool in the diagnosis of Alzheimer's disease: A review of the evidence. Neuropsychology Review. 2011; 21(1):54–65. [PubMed: 21086049]

*. Chandler MJ, Lacritz LH, Hynan LS, Barnard HD, Allen G, Deschner M, Weiner MF, Cullum CM. A total score for the CERAD neuropsychological battery. Neurology. 2005; 65(1):102–106. [PubMed: 16009893]

Chapman KR, Bing-Canar H, Alosco ML, Steinberg EG, Martin B, Chaisson C, et al. Mini mental state examination and logical memory scores for entry into Alzheimer's disease trials. Alzheimer's Research & Therapy. 2016; 8:9.

Chelune GJ. Evidence-based research and practice in clinical neuropsychology. The Clinical Neuropsychologist. 2010; 24(3):454–467. [PubMed: 18821179]

*. Clark JH, Hobson VL, O'Bryant SE. Diagnostic accuracy of percent retention scores on RBANS verbal memory subtests for the diagnosis of Alzheimer's disease and mild cognitive impairment. Archives of Clinical Neuropsychology. 2010; 25(4):318–326. [PubMed: 20378680]

Clark LR, Delano-Wood L, Libon DJ, McDonald CR, Nation DA, Bangen KJ, et al. Are empirically-derived subtypes of mild cognitive impairment consistent with conventional subtypes? Journal of the International Neuropsychological Society. 2013; 19(06):635–645. [PubMed: 23552486]

*. Crocco E, Curiel RE, Acevedo A, Czaja SJ, Loewenstein DA. An evaluation of deficits in semantic cueing and proactive and retroactive interference as early features of Alzheimer's disease. The American Journal of Geriatric Psychiatry. 2014; 22(9):889–897. [PubMed: 23768680]

Davis D, Schmitt F, Wekstein D, Markesbery W. Alzheimer neuropathologic alterations in aged cognitively normal subjects. Journal of Neuropathology & Experimental Neurology. 1999; 58(4): 376–388. [PubMed: 10218633]

*. de Jager CA, Hogervorst E, Combrinck M, Budge MM. Sensitivity and specificity of neuropsychological tests for mild cognitive impairment, vascular cognitive impairment and Alzheimer's disease. Psychological Medicine. 2003; 33(6):1039–1050. [PubMed: 12946088]

*. Delgado C, Munoz-Neira C, Soto A, Martinez M, Henriquez F, Flores P, et al. Comparison of the psychometric properties of the "word" and "picture" versions of the free and cued selective reminding test in a Spanish-speaking cohort of patients with mild Alzheimer's disease and

cognitively healthy controls. Archives of Clinical Neuropsychology. 2016; 31(2):165–175. [PubMed: 26758367]

Delis DC, Massman PJ, Butters N, Salmon DP, Cermak LS, Kramer JH. Profiles of demented and amnesic patients on the California verbal learning test: Implications for the assessment of memory disorders. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1991; 3(1):19.

Delis DC, Wetter SR, Jacobson MW, Peavy G, Hamilton J, Gongvatana A, et al. Recall discriminability: Utility of a new CVLT-II measure in the differential diagnosis of dementia. Journal of the International Neuropsychological Society. 2005; 11(6):708–715. [PubMed: 16248906]

*. Delprado J, Kinsella G, Ong B, Pike K, Ames D, Storey E, et al. Clinical measures of prospective memory in amnestic mild cognitive impairment. Journal of the International Neuropsychological Society. 2012; 18(2):295–304. [PubMed: 22264396]

Dennis, ML., Lennox, MA., Foss, MA. Practical power analysis for substance abuse health services research. In: Bryant, K.Windle, M., West, SG., editors. The science of prevention: Methodological advances from alcohol and substance abuse research. Washington DC: American Psychological Association; 1997. p. 367-404.

DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7(3):177–188. [PubMed: 3802833]

Devanand DP, Liu X, Tabert MH, Pradhaban G, Cuasay K, Bell K, et al. Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. Biological Psychiatry. 2008; 64(10):871–879. [PubMed: 18723162]

Doebler, P. MADA: Meta-analysis of diagnostic accuracy. 2015. Available from: cran.r-project.org/web/packages/mada/vignettes/mada.pdf. R package version 0.5. 7

Doebler, P., Holling, H. Meta-analysis of diagnostic accuracy with mada. 2015. Retrieved at: https://cran.rproject.org/web/packages/mada/vignettes/mada.pdf

Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria. Lancet Neurology. 2007; 6(8):734–746. [PubMed: 17616482]

Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, et al. Revising the definition of Alzheimer's disease: A new lexicon. Lancet Neurology. 2010; 9(11):1118–1127. [PubMed: 20934914]

Ducharme S, Guiot M-C, Nikelski J, Chertkow H. Does a positive Pittsburgh compound B scan in a patient with dementia equal Alzheimer disease? JAMA Neurology. 2013; 70(7):912–914. [PubMed: 23689280]

*. Duff K, Humphreys Clark JD, O'Bryant SE, Mold JW, Schiffer RB, Sutker PB. Utility of the RBANS in detecting cognitive impairment associated with Alzheimer's disease: Sensitivity, specificity, and positive and negative predictive powers. Archives of Clinical Neuropsychology. 2008; 23(5):603–612. [PubMed: 18639437]

*. Duff K, Hobson VL, Beglinger LJ, O'Bryant SE. Diagnostic accuracy of the RBANS in mild cognitive impairment: Limitations on assessing milder impairments. Archives of Clinical Neuropsychology. 2010; 25(5):429–441. [PubMed: 20570820]

Eckerström C, Olsson E, Bjerke M, Malmgren H, Edman Å, Wallin A, et al. A combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts conversion from mild cognitive impairment to dementia. Journal of Alzheimer's Disease. 2013; 36(3):421–431.

Edison P, Rowe CC, Rinne JO, Ng S, Ahmed I, Kemppainen N, et al. Amyloid load in Parkinson's disease dementia and Lewy body dementia measured with [11C] PIB positron emission tomography. Journal of Neurology, Neurosurgery & Psychiatry. 2008; 79(12):1331–1338.

Edmonds EC, Delano-Wood L, Clark LR, Jak AJ, Nation DA, McDonald CR, et al. Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. Alzheimer's & Dementia. 2015a; 11:415–424.

Edmonds EC, Delano-Wood L, Galasko D, Salmon DP, Bondi MW. Subtle cognitive decline and biomarker staging in preclinical Alzheimer's disease. Journal of Alzheimer's Disease. 2015b; 47:231–242.

Ellis, PD. The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge: Cambridge University press; 2010.

*. Ewers M, Walsh C, Trojanowski JQ, Shaw LM, Petersen RC, Jack CR Jr, et al. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiology of Aging. 2012; 33(7):1203–1214. [PubMed: 21159408]

Fields JA, Ferman TJ, Boeve BF, Smith GE. Neuropsychological assessment of patients with dementing illness. Nature Reviews. Neurology. 2011; 7(12):677–687. [PubMed: 22045270]

Fischer P, Jungwirth S, Zehetmayer S, Weissgram S, Hoenigschnabl S, Gelpi E, et al. Conversion from subtypes of mild cognitive impairment to Alzheimer dementia. Neurology. 2007; 68(4):288–291. [PubMed: 17242334]

Gainotti G, Quaranta D, Vita MG, Marra C. Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease. Journal of Alzheimer's Disease. 2014; 38(3):481–495.

Ganguli M, Snitz BE, Saxton JA, Chang C-CH, Lee C-W, Vander Bilt J, et al. Outcomes of mild cognitive impairment by definition: A population study. Archives of Neurology. 2011; 68(6):761–767. [PubMed: 21670400]

*. Gavett BE, Poon SJ, Ozonoff A, Jefferson AL, Nair AK, Green RC, et al. Diagnostic utility of the NAB list learning test in Alzheimer's disease and amnestic mild cognitive impairment. Journal of the International Neuropsychological Society. 2009; 15(1):121–129. [PubMed: 19128535]

Gavett BE, Lou KR, Daneshvar DH, Green RC, Jefferson AL, Stern RA. Diagnostic accuracy statistics for seven neuropsychological assessment battery (NAB) test variables in the diagnosis of Alzheimer's disease. Appl Neuropsychol Adult. 2012; 19(2):108–115. [PubMed: 23373577]

Gomar JJ, Conejero-Goldberg C, Davies P, Goldberg TE, Initiative AsDN. Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data. Alzheimer's & Dementia. 2014; 10(6):704–712.

*. Gonzalez-Palau F, Franco M, Jimenez F, Parra E, Bernate M, Solis A. Clinical utility of the hopkins verbal test-revised for detecting Alzheimer's disease and mild cognitive impairment in Spanish population. Archives of Clinical Neuropsychology. 2013; 28(3):245–253. [PubMed: 23384601]

Graham NL, Emery T, Hodges JR. Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia. Journal of Neurology, Neurosurgery, and Psychiatry. 2004; 75(1):61–71.

Grober E, Buschke H. Genuine memory deficits in dementia. Developmental Neuropsychology. 1987; 3(1):13–36.

Growdon J, Selkoe DJ, Roses AD, et al. The Ronald and Nancy Reagan research Institute of the Alzheimer’s association and the National Institute on Aging working group. Consensus report of the working group on: molecular and biochemical markers of Alzheimer’s disease. Neurobiology of Aging. 1998; 19(2):109–116. [PubMed: 9558143]

Han JW, Kim TH, Lee SB, Park JH, Lee JJ, Huh Y, et al. Predictive validity and diagnostic stability of mild cognitive impairment subtypes. Alzheimers Dement. 2012; 8(6):553–559. [PubMed: 23102125]

Han, SD., Nguyen, CP., Stricker, NH., Nation, DA. Detectable neuropsychological differences in early preclinical Alzheimer's disease: A meta-analysis. Neuropsychology Review. 2017. https://doi.org/10.1007/s11065-017-9345-5

Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. Science. 2002; 297(5580):353–356. [PubMed: 12130773]

Heister D, Brewer JB, Magda S, Blennow K, McEvoy LK, Initiative AsDN. Predicting MCI outcome with clinically available MRI and CSF biomarkers. Neurology. 2011; 77(17):1619–1628. [PubMed: 21998317]

Herrup K. The case for rejecting the amyloid cascade hypothesis. Nature Neuroscience. 2015; 18:794–799. [PubMed: 26007212]

*. Hogervorst E, Combrinck M, Lapuerta P, Rue J, Swales K, Budge M. The Hopkins verbal learning test and screening for dementia. Dementia and Geriatric Cognitive Disorders. 2002; 13(1):13–20. [PubMed: 11731710]

*. Ivanoiu A, Adam S, Linden M, Salmon E, Juillerat A-C, Mulligan R, Seron X. Memory evaluation with a new cued recall test in patients with mild cognitive impairment and Alzheimer's disease. Journal of Neurology. 2005; 252(1):47–55. [PubMed: 15654553]

*. Ivnik RJ, Smith GE, Petersen RC, Boeve BF, Kokmen E, Tangalos EG. Diagnostic accuracy of four approaches to interpreting neuropsychological test data. Neuropsychology. 2000; 14(2):163–177. [PubMed: 10791857]

Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011; 7(3):257–262.

Jack CR, Bennett DA, Blennow K, Carrillo MC, Feldman HH, Giovanni B, et al. A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. Neurology. 2016; 87(5):539–547. [PubMed: 27371494]

Jak AJ, Bondi MW, Delano-Wood L, Wierenga C, Corey-Bloom J, Salmon DP, et al. Quantification of five neuropsychological approaches to defining mild cognitive impairment. The American Journal of Geriatric Psychiatry. 2009; 17(5):368–375. [PubMed: 19390294]

Johnson KA, Minoshima S, Bohnen NI, Donohoe KJ, Foster NL, Herscovitch P, et al. Appropriate use criteria for amyloid PET: A report of the amyloid imaging task force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer’s association. Journal of Nuclear Medicine. 2013a; 54(3):476–490. [PubMed: 23359661]

Johnson KA, Minoshima S, Bohnen NI, Donohoe KJ, Foster NL, Herscovitch P, et al. Appropriate use criteria for amyloid PET: A report of the amyloid imaging task force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association. Alzheimers Dement. 2013b; 9(1):e-1–e16. [PubMed: 22402324]

*. Junkkila J, Oja S, Laine M, Karrasch M. Applicability of the CANTAB-PAL computerized memory test in identifying amnestic mild cognitive impairment and Alzheimer's disease. Dementia and Geriatric Cognitive Disorders. 2012; 34(2):83–89. [PubMed: 22922741]

Kanne SM, Balota DA, Storandt M, McKeel DW Jr, Morris JC. Relating anatomy to function in Alzheimer's disease: Neuropsychological profiles predict regional neuropathology 5 years later. Neurology. 1998; 50(4):979–985. [PubMed: 9566382]

*. Karantzoulis S, Novitski J, Gold M, Randolph C. The repeatable battery for the assessment of neuropsychological status (RBANS): Utility in detection and characterization of mild cognitive impairment due to Alzheimer's disease. Archives of Clinical Neuropsychology. 2013; 28(8):837–844. [PubMed: 23867976]

Karrasch M, Sinerva E, Gronholm P, Rinne J, Laine M. CERAD test performances in amnestic mild cognitive impairment and Alzheimer's disease. Acta Neurologica Scandinavica. 2005; 111(3):172–179. [PubMed: 15691286]

Kim KW, Lee J, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers-part I. General guidance and tips. Korean Journal of Radiology. 2015; 16(6):1175–1187. [PubMed: 26576106]

Klekociuk S, Summers J, Vickers J, Summers MJ. Reducing false positive diagnoses in mild cognitive impairment: The importance of comprehensive neuropsychological assessment. European Journal of Neurology. 2014; 21(10):1330–e1383. [PubMed: 24943259]

Laforce R, Rabinovici GD. Amyloid imaging in the differential diagnosis of dementia: Review and potential clinical applications. Alzheimer's Research & Therapy. 2011; 3(6):1.

*. Kuslansky G, Katz M, Verghese J, Hall CB, Lapuerta P, LaRuffa G, Lipton RB. Detecting dementia with the Hopkins Verbal Learning Test and the Mini-Mental State Examination. Archives of Clinical Neuropsychology. 2004; 19(1):89–104. [PubMed: 14670382]

Landau S, Harvey D, Madison C, Reiman E, Foster N, Aisen P, et al. Comparing predictors of conversion and decline in mild cognitive impairment. Neurology. 2010; 75(3):230–238. [PubMed: 20592257]

Lee J, Kim KW, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers-part II. Statistical methods of meta-analysis. Korean Journal of Radiology. 2015; 16(6):1188–1196. [PubMed: 26576107]

*. Lehrner J, Maly J, Gleiß A, Auff E, Dal-Bianco P. Demenzdiagnostik mit Hilfe der Vienna Neuropsychologischen Testbatterie (VNTB): Standardisierung, Normierung und Validierung. Psychologie in Österreich. 2007; 4&5:358–365.

*. Lekeu F, Magis D, Marique P, Delbeuck X, Bechet S, Guillaume B, et al. The California verbal learning test and other standard clinical neuropsychological tests to predict conversion from mild memory impairment to dementia. Journal of Clinical and Experimental Neuropsychology. 2010; 32(2):164–173. [PubMed: 19459119]

*. Lemos R, Duro D, Simões MR, Santana I. The free and cued selective reminding test distinguishes frontotemporal dementia from Alzheimer's disease. Archives of Clinical Neuropsychology. 2014; 29(7):670–679. [PubMed: 25062746]

*. Lemos R, Simoes MR, Santiago B, Santana I. The free and cued selective reminding test: Validation for mild cognitive impairment and Alzheimer's disease. Journal of Neuropsychology. 2015; 9(2): 242–257. [PubMed: 24894485]

Lin, JS., O'Connor, E., Rossom, RC., Perdue, LA., Burda, BU., Thompson, M., et al. Screening for Cognitive Impairment in Older Adults: An Evidence Update for the U.S. Preventive Services Task Force. Agency for Healthcare Research and Quality (US); Rockville, Md: 2013. U.S. Preventive Services Task Force evidence syntheses, formerly systematic evidence reviews.

Lindeboom J, Schmand B, Tulner L, Walstra G, Jonker C. Visual association test to detect early dementia of the Alzheimer type. Journal of Neurology, Neurosurgery, and Psychiatry. 2002; 73(2): 126–133.

Lineweaver TT, Bondi MW, Galasko D, Salmon DP. Effect of knowledge of APOE genotype on subjective and objective memory performance in healthy older adults. American Journal of Psychiatry. 2014; 171(2):201–208. [PubMed: 24170170]

*. Loewenstein DA, Acevedo A, Luis C, Crum T, Barker WW, Duara R. Semantic interference deficits and the detection of mild Alzheimer's disease and mild cognitive impairment without dementia. Journal of the International Neuropsychological Society. 2004; 10(1):91–100. [PubMed: 14751011]

*. Loewenstein DA, Acevedo A, Ownby R, Agron J, Barker WW, Isaacson R, et al. Using different memory cutoffs to assess mild cognitive impairment. The American Journal of Geriatric Psychiatry. 2006; 14(11):911–919. [PubMed: 17068313]

Looi JC, Sachdev PS. Differentiation of vascular dementia from AD on neuropsychological tests. Neurology. 1999; 53(4):670–678. [PubMed: 10489025]

Lowndes G, Savage G. Early detection of memory impairment in Alzheimer’s disease: A neurocognitive perspective on assessment. Neuropsychology Review. 2007; 17(3):193–202. [PubMed: 17805975]

Lowndes GJ, Saling MM, Ames D, Chiu E, Gonzalez LM, Savage GR. Recall and recognition of verbal paired associates in early Alzheimer's disease. Journal of the International Neuropsychological Society. 2008; 14(4):591–600. [PubMed: 18577288]

Lundervold AJ, Reinvang I, Lundervold A. Characteristic patterns of verbal memory function in patients with Huntington's disease. Scandinavian Journal of Psychology. 1994; 35(1):38–47. [PubMed: 8191260]

Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 1.0. 2009

Marcone S, Gagnon JF, Lecomte S, Imbeault H, Limoges F, Postuma RB, et al. Clinical utility of the envelope task in mild cognitive impairment and dementia. The Canadian Journal of Neurological Sciences. 2017; 44(1):9–16. [PubMed: 27665668]

Matioli MNP, Caramelli P. Limitations in differentiating vascular dementia from Alzheimer's disease with brief cognitive tests. Arquivos de Neuro-Psiquiatria. 2010; 68(2):185–188. [PubMed: 20464282]

McDermott AT, DeFilippis NA. Are the indices of the RBANS sufficient for differentiating Alzheimer's disease and subcortical vascular dementia? Archives of Clinical Neuropsychology. 2010; 25(4):327–334. [PubMed: 20430863]

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 the Department of Health and Human Services Task Force on Alzheimer's disease. Neurology. 1984; 34:939–944. [PubMed: 6610841]

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, 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. Alzheimers Dement. 2011; 7(3):263–269. [PubMed: 21514250]

Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. Acta Psychiatrica Scandinavica. 2009; 119(4): 252–265. [PubMed: 19236314]

Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Annals of Internal Medicine. 2009; 151(4):264–269. [PubMed: 19622511]

Naslund J, Haroutunian V, Mohs R, Davis KL, Davies P, Greengard P, et al. Correlation between elevated levels of amyloid beta-peptide in the brain and cognitive decline. JAMA. 2000; 283(12): 1571–1577. [PubMed: 10735393]

Noel-Storr AH, McCleery JM, Richard E, Ritchie CW, Flicker L, Cullum SJ, et al. Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem initiative. Neurology. 2014; 83(4):364–373. [PubMed: 24944261]

O'Carroll R, Conway S, Ryman A, Prentice N. Performance on the delayed word recall test (DWR) fails to differentiate clearly between depression and Alzheimer's disease in the elderly. Psychological Medicine. 1997; 27(4):967–971. [PubMed: 9234474]

O'Connell ME, Tuokko H. Age corrections and dementia classification accuracy. Archives of Clinical Neuropsychology. 2010; 25(2):126–138. [PubMed: 20118110]

O'Connell H, Coen R, Kidd N, Warsi M, Chin AV, Lawlor BA. Early detection of Alzheimer's disease (AD) using the CANTAB paired associates learning test. International Journal of Geriatric Psychiatry. 2004; 19(12):1207–1208. [PubMed: 15578796]

Oda H, Yamamoto Y, Maeda K. The neuropsychological profile in dementia with Lewy bodies and Alzheimer's disease. International Journal of Geriatric Psychiatry. 2009; 24(2):125–131. [PubMed: 18615776]

Palmqvist S, Hertze J, Minthon L, Wattmo C, Zetterberg H, Blennow K, et al. Comparison of brief cognitive tests and CSF biomarkers in predicting Alzheimer’s disease in mild cognitive impairment: Six-year follow-up study. PloS One. 2012; 7(6):e38639. [PubMed: 22761691]

*. Park S, Kim I, Park HG, Shin SA, Cho Y, Youn JH, et al. Development and validation of the rappel Indice-24: Behavioral and brain morphological evidence. Journal of Geriatric Psychiatry and Neurology. 2016; 29(3):160–168. [PubMed: 26956224]

*. Parra MA, Abrahams S, Logie RH, Mendez LG, Lopera F, Della Sala S. Visual short-term memory binding deficits in familial Alzheimer's disease. Brain. 2010; 133(9):2702–2713. [PubMed: 20624814]

Parra MA, Sala SD, Abrahams S, Logie RH, Mendez LG, Lopera F. Specific deficit of colour-colour short-term memory binding in sporadic and familial Alzheimer's disease. Neuropsychologia. 2011; 49(7):1943–1952. [PubMed: 21435348]

*. Parra MA, Ascencio LL, Urquina HF, Manes F, Ibanez AM. P300 and neuropsychological assessment in mild cognitive impairment and Alzheimer dementia. Frontiers in Neurology. 2012; 3:172. [PubMed: 23227021]

*. Paulsen JS, Salmon DP, Monsch AU, Butters N, Swenson MR, Bondi MW. Discrimination of cortical from subcortical dementias on the basis of memory and problem-solving tests. Journal of Clinical Psychology. 1995; 51(1):48–58. [PubMed: 7782475]

*. Pengas G, Patterson K, Arnold RJ, Bird CM, Burgess N, Nestor PJ. Lost and found: Bespoke memory testing for Alzheimer's disease and semantic dementia. Journal of Alzheimer's Disease. 2010; 21(4):1347–1365.

Perani D, Schillaci O, Padovani A, Nobili F, Iaccarino L, Della Rosa P, et al. A survey of FDG-and amyloid-PET imaging in dementia and GRADE analysis. BioMed Research International. 2013; 2014:785039–785039.

Peters F, Villeneuve S, Belleville S. Predicting progression to dementia in elderly subjects with mild cognitive impairment using both cognitive and neuroimaging predictors. Journal of Alzheimer's Disease. 2014; 38(2):307–318.

Petersen RC. Mild cognitive impairment as a diagnostic entity. Journal of Internal Medicine. 2004; 256(3):183–194. [PubMed: 15324362]

Petersen R. Mild cognitive impairment. Continuum: Lifelong learning in neurology. Dementia. 2007; 13(2):15–38.

Petersen R, Kanow C. Mild cognitive impairment-state of the art 2001. REVUE NEUROLOGIQUE. 2001; 157(10; SUPP):4S29–24S29.

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):303–308. [PubMed: 10190820]

Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, et al. Current concepts in mild cognitive impairment. Archives of Neurology. 2001; 58(12):1985–1992. [PubMed: 11735772]

Petersen RC, Roberts RO, Knopman DS, Boeve BF, Geda YE, Ivnik RJ, et al. Mild cognitive impairment: Ten years later. Archives of Neurology. 2009; 66(12):1447–1455. [PubMed: 20008648]

Petersen RC, Aisen P, Boeve BF, Geda YE, Ivnik RJ, Knopman DS, et al. Mild cognitive impairment due to Alzheimer disease in the community. Annals of Neurology. 2013; 74(2):199–208. [PubMed: 23686697]

*. Pike KE, Kinsella GJ, Ong B, Mullaly E, Rand E, Storey E, et al. Is the WMS-IV verbal paired associates as effective as other memory tasks in discriminating amnestic mild cognitive impairment from normal aging? The Clinical Neuropsychologist. 2013; 27(6):908–923. [PubMed: 23767765]

Portet F, Ousset PJ, Visser PJ, Frisoni GB, Nobili F, Scheltens P, et al. Mild cognitive impairment (MCI) in medical practice: A critical review of the concept and new diagnostic procedure. Report of the MCI working Group of the European Consortium on Alzheimer's disease. Journal of Neurology, Neurosurgery, and Psychiatry. 2006; 77(6):714–718.

*. Rabin LA, Pare N, Saykin AJ, Brown MJ, Wishart HA, Flashman LA, et al. Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer's disease. Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition. 2009; 16(3):357–376.

Ravaglia G, Forti P, Maioli F, Martelli M, Servadei L, Brunetti N, et al. Conversion of mild cognitive impairment to dementia: Predictive role of mild cognitive impairment subtypes and vascular risk factors. Dementia and Geriatric Cognitive Disorders. 2005; 21(1):51–58. [PubMed: 16276110]

Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. Cochrane handbook for systematic reviews of diagnostic test accuracy version. 2009; 1(0):1–27.

Rentz DM, Parra Rodriguez MA, Amariglio R, Stern Y, Sperling R, Ferris S. Promising developments in neuropsychological approaches for the detection of preclinical Alzheimer's disease: A selective review. Alzheimer's Research & Therapy. 2013; 5(6):58.

*. Ritter E, Despres O, Monsch AU, Manning L. Topographical recognition memory sensitive to amnestic mild cognitive impairment but not to depression. International Journal of Geriatric Psychiatry. 2006; 21(10):924–929. [PubMed: 16927398]

Saka E, Elibol B. Enhanced cued recall and clock drawing test performances differ in Parkinson's and Alzheimer's disease-related cognitive dysfunction. Parkinsonism & Related Disorders. 2009; 15(9):688–691. [PubMed: 19446489]

Saka E, Mihci E, Topcuoglu MA, Balkan S. Enhanced cued recall has a high utility as a screening test in the diagnosis of Alzheimer's disease and mild cognitive impairment in Turkish people. Archives of Clinical Neuropsychology. 2006; 21(7):745–751. [PubMed: 16979317]

Salmon DP, Granholm E, McCullough D, Butters N, Grant I. Recognition memory span in mildly and moderately demented patients with Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology. 1989; 11(4):429–443. [PubMed: 2760179]

*. Salmon DP, Thomas RG, Pay MM, Booth A, Hofstetter CR, Thal LJ, et al. Alzheimer's disease can be accurately diagnosed in very mildly impaired individuals. Neurology. 2002; 59(7):1022–1028. [PubMed: 12370456]

Schmidtke K, Hüll M. Neuropsychological differentiation of small vessel disease, Alzheimer's disease and mixed dementia. Journal of the Neurological Sciences. 2002; 203:17–22. [PubMed: 12417351]

*. Schrijnemaekers AM, de Jager CA, Hogervorst E, Budge MM. Cases with mild cognitive impairment and Alzheimer's disease fail to benefit from repeated exposure to episodic memory tests as compared with controls. Journal of Clinical and Experimental Neuropsychology. 2006; 28(3):438–455. [PubMed: 16618630]

Schneider JA, Arvanitakis Z, Leurgans SE, Bennett DA. The neuropathology of probable Alzheimer disease and mild cognitive impairment. Annals of Neurology. 2009; 66(2):200–208. [PubMed: 19743450]

Schwarzer, G., Carpenter, JR., Rücker, G. Meta-analysis with R. New York, NY: Springer; 2015.

*. Serna A, Contador I, Bermejo-Pareja F, Mitchell AJ, Fernández-Calvo B, Ramos F, et al. Accuracy of a brief neuropsychological battery for the diagnosis of dementia and mild cognitive impairment: An analysis of the NEDICES cohort. Journal of Alzheimer's Disease. 2015; 48:163–173.

*. Shankle WR, Romney AK, Hara J, Fortier D, Dick MB, Chen JM, et al. Methods to improve the detection of mild cognitive impairment. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(13):4919–4924. [PubMed: 15781874]

*. Shi J, Tian J, Wei M, Miao Y, Wang Y. The utility of the Hopkins verbal learning test (Chinese version) for screening dementia and mild cognitive impairment in a Chinese population. BMC Neurology. 2012; 12:136. [PubMed: 23130844]

*. Shi J, Wei M, Tian J, Snowden J, Zhang X, Ni J, et al. The Chinese version of story recall: A useful screening tool for mild cognitive impairment and Alzheimer's disease in the elderly. Biomed Central Psychiatry. 2014; 14(71):1–10.

Smith, GE., Ivnik, RI., Lucas, J. Assessment techniques: Tests, test batteries, norms, and methodological approaches. In: Morgan, JE., Ricker, JH., editors. Textbook of clinical neuropsychology. New York: Taylor & Francis Group; 2008. p. 38-58.

*. Sotaniemi M, Pulliainen V, Hokkanen L, Pirttila T, Hallikainen I, Soininen H, et al. CERAD-neuropsychological battery in screening mild Alzheimer's disease. Acta Neurologica Scandinavica. 2012; 125(1):16–23. [PubMed: 21198445]

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011; 7(3):280–292.

*. Storandt M, Morris JC. Ascertainment bias in the clinical diagnosis of Alzheimer disease. Archives of Neurology. 2010; 67(11):1364–1369. [PubMed: 21060013]

Summers MJ, Saunders NL. Neuropsychological measures predict decline to Alzheimer's dementia from mild cognitive impairment. Neuropsychology. 2012; 26(4):498. [PubMed: 22612573]

*. Takechi H, Dodge HH. Scenery picture memory test: A new type of quick and effective screening test to detect early stage Alzheimer's disease patients. Geriatrics & Gerontology International. 2010; 10(2):183–190. [PubMed: 20446933]

*. Testa JA, Ivnik RJ, Boeve B, Petersen RC, Pankratz VS, Knopman D, et al. Confrontation naming does not add incremental diagnostic utility in MCI and Alzheimer's disease. Journal of the International Neuropsychological Society. 2004; 10(4):504–512. [PubMed: 15327729]

Tierney MC, Black SE, Szalai JP, Snow WG, Fisher RH, Nadon G, et al. Recognition memory and verbal fluency differentiate probable Alzheimer disease from subcortical ischemic vascular dementia. Archives of Neurology. 2001; 58(10):1654–1659. [PubMed: 11594925]

Troster AI, Butters N, Salmon DP, Cullum CM, Jacobs D, Brandt J, et al. The diagnostic utility of savings scores: Differentiating Alzheimer's and Huntington's diseases with the logical memory and visual reproduction tests. Journal of Clinical and Experimental Neuropsychology. 1993; 15(5):773–788. [PubMed: 8276935]

*. Thompson TAC, Wilson PH, Snyder PJ, Pietrzak RH, Darby D, Maruff P, Buschke H. Sensitivity and Test-Retest Reliability of the International Shopping List Test in Assessing Verbal Learning and Memory in Mild Alzheimer's Disease. Archives of Clinical Neuropsychology. 2011; 26(5): 412–424. [PubMed: 21613302]

Troyer AK, Murphy KJ, Anderson ND, Hayman-Abello BA, Craik FI, Moscovitch M. Item and associative memory in amnestic mild cognitive impairment: Performance on standardized memory tests. Neuropsychology. 2008; 22(1):10–16. [PubMed: 18211151]

Troyer AK, Murphy KJ, Anderson ND, Craik FI, Moscovitch M, Maione A, et al. Associative recognition in mild cognitive impairment: Relationship to hippocampal volume and apolipoprotein E. Neuropsychologia. 2012; 50(14):3721–3728. [PubMed: 23103838]

Turriziani P, Fadda L, Caltagirone C, Carlesimo GA. Recognition memory for single items and for associations in amnesic patients. Neuropsychologia. 2004; 42(4):426–433. [PubMed: 14728917]

US Dept of Health and Human Services, US Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for industry. Alzheimer’s disease: developing drugs for the treatment of early stage disease (draft guidance). 2013. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM338287.pdf

Vacante M, Wilcock GK, de Jager CA. Computerized adaptation of the placing test for early detection of both mild cognitive impairment and Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology. 2013; 35(8):846–856. [PubMed: 23985007]

*. Vogel A, Mortensen EL, Gade A, Waldemar G. The category cued recall test in very mild Alzheimer's disease: Discriminative validity and correlation with semantic memory functions. European Journal of Neurology. 2007; 14(1):102–108. [PubMed: 17222122]

Wang HM, Yang CM, Kuo WC, Huang CC, Kuo HC. Use of a modified spatial-context memory test to detect amnestic mild cognitive impairment. PloS One. 2013; 8(2):e57030. [PubMed: 23468906]

Wechsler D. A Standardized Memory Scale for Clinical Use. The Journal of Psychology. 1945; 19(1): 87–95. https://doi.org/10.1080/00223980.1945.9917223.

Wechsler, D. WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale. Psychological corporation; San Antonio, TX: 1997.

*. Welsh K, Butters N, Hughes J, Mohs R, Heyman A. Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. Archives of Neurology. 1991; 48(3):278–281. [PubMed: 2001185]

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):529–536. [PubMed: 22007046]

Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund LO, 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):240–246. [PubMed: 15324367]

Yaffe K, Petersen RC, Lindquist K, Kramer J, Miller B. Subtype of mild cognitive impairment and progression to dementia and death. Dementia and Geriatric Cognitive Disorders. 2006; 22(4): 312–319. [PubMed: 16940725]

Yassuda MS, Flaks MK, Viola LF, Pereira FS, Memoria CM, Nunes PV, et al. Psychometric characteristics of the Rivermead Behavioural memory test (RBMT) as an early detection instrument for dementia and mild cognitive impairment in Brazil. International Psychogeriatrics. 2010; 22(6):1003–1011. [PubMed: 20598195]

Yonelinas AP, Hopfinger JB, Buonocore MH, Kroll NE, Baynes K. Hippocampal, parahippocampal and occipital-temporal contributions to associative and item recognition memory: An fMRI study. Neuroreport. 2001; 12(2):359–363. [PubMed: 11209950]

Zhang S, Smailagic N, Hyde C, Noel-Storr AH, Takwoingi Y, McShane R, et al. 11C–PIB-PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev. 2014; 7:CD010386.

Zlokovic BV. Neurovascular pathways to neurodegeneration in Alzheimer's disease and other disorders. Nature Reviews Neuroscience. 2011; 12(12):723–738. [PubMed: 22048062]

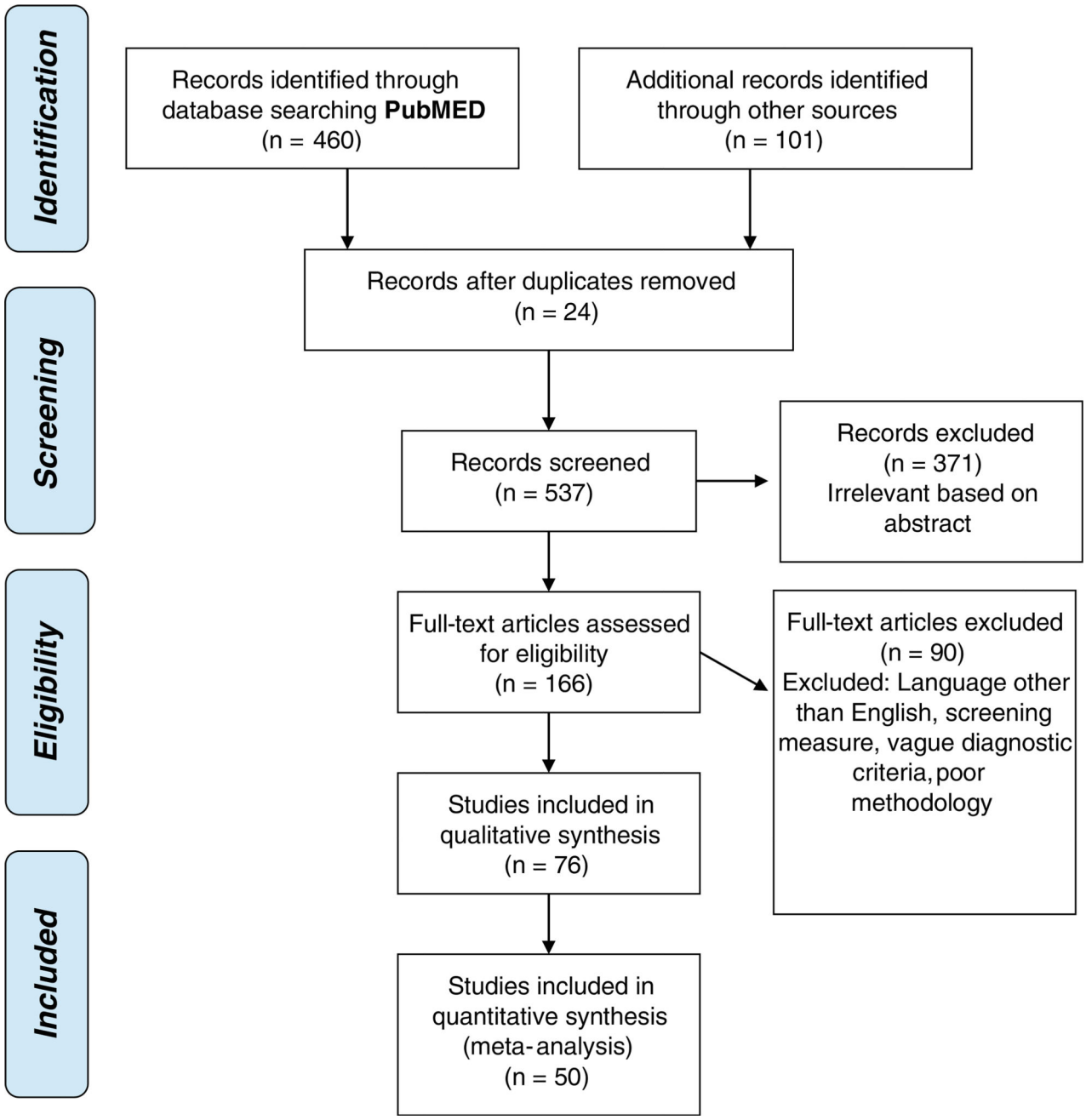


**Fig. 1.**

Final number of studies meeting inclusionary criteria based on PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) 2009 standards

| | Study | AD | HC | TP | FP | FN | TN | Sensitivity (95%, CI) | Sensitivity (95%, CI) | Specificity (95%, CI) | Specificity (95%, CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Adam et al. (2007) | 38 | 57 | 35 | 3 | 3 | 54 | | 0.92 [0.79, 0.97] | | 0.95 [0.86, 0.98] |
| 2 | Baek et al. (2012) * | 72 | 53 | 58 | 14 | 14 | 39 | | 0.81 [0.70, 0.88] | | 0.74 [0.60, 0.84] |
| 3 | Bertolucci et al. (2001) | 43 | 85 | 37 | 11 | 6 | 74 | | 0.86 [0.73, 0.93] | | 0.87 [0.78, 0.93] |
| 4 | Cahn et al. (1995) | 40 | 236 | 36 | 50 | 4 | 186 | | 0.90 [0.77, 0.96] | | 0.79 [0.73, 0.84] |
| 5 | de Jager et al. (2003) * | 60 | 51 | 56 | 4 | 4 | 47 | | 0.93 [0.84, 0.97] | | 0.92 [0.82, 0.97] |
| 6 | Duff et al. (2008) * | 69 | 69 | 61 | 11 | 8 | 58 | | 0.88 [0.79, 0.94] | | 0.84 [0.74, 0.91] |
| 7 | Ewers et al. (2012) | 81 | 101 | 73 | 9 | 8 | 92 | | 0.90 [0.82, 0.95] | | 0.91 [0.84, 0.95] |
| 8 | Gavett et al. (2009) | 26 | 98 | 22 | 10 | 4 | 88 | | 0.85 [0.66, 0.94] | | 0.90 [0.82, 0.94] |
| 9 | Gonzalez-Palau et al. (2013) | 54 | 109 | 46 | 13 | 8 | 96 | | 0.85 [0.73, 0.92] | | 0.88 [0.81, 0.93] |
| 10 | Hogervorst et al. (2002) | 82 | 114 | 75 | 2 | 7 | 112 | | 0.91 [0.83, 0.96] | | 0.98 [0.94, 1.00] |
| 11 | Ivanoiu et al. (2005) * | 22 | 22 | 20 | 1 | 2 | 21 | | 0.91 [0.72, 0.97] | | 0.95 [0.78, 0.99] |
| 12 | Ivnik et al. (2000) | 232 | 432 | 195 | 69 | 37 | 363 | | 0.84 [0.79, 0.88] | | 0.84 [0.80, 0.87] |
| 13 | Kuslansky et al. (2004) | 70 | 323 | 58 | 55 | 12 | 268 | | 0.83 [0.72, 0.90] | | 0.83 [0.78, 0.87] |
| 14 | Lehrner et al. (2007) | 78 | 58 | 71 | 4 | 7 | 54 | | 0.91 [0.83, 0.96] | | 0.93 [0.84, 0.97] |
| 15 | Lemos et al. (2014) | 70 | 101 | 66 | 1 | 4 | 100 | | 0.94 [0.86, 0.98] | | 0.99 [0.95, 1.00] |
| 16 | Park et al. (2016) | 28 | 34 | 21 | 3 | 7 | 31 | | 0.75 [0.57, 0.87] | | 0.91 [0.77, 0.97] |
| 17 | Parra et al (2012) | 10 | 10 | 9 | 2 | 1 | 8 | | 0.90 [0.60, 0.98] | | 0.80 [0.49, 0.94] |
| 18 | Paulsen et al. (1995) | 20 | 20 | 13 | 4 | 7 | 16 | | 0.65 [0.43, 0.82] | | 0.80 [0.58, 0.92] |
| 19 | Pengas et al. (2010) | 22 | 33 | 21 | 2 | 1 | 31 | | 0.95 [0.78, 0.99] | | 0.94 [0.80, 0.98] |
| 20 | Salmon et al. (2002) | 98 | 98 | 93 | 11 | 5 | 87 | | 0.95 [0.89, 0.98] | | 0.89 [0.81, 0.94] |
| 21 | Shi et al. (2012) | 97 | 249 | 92 | 17 | 5 | 232 | | 0.95 [0.88, 0.98] | | 0.93 [0.89, 0.96] |
| 22 | Sotaniemi et al. (2012) | 171 | 315 | 154 | 54 | 17 | 261 | | 0.90 [0.85, 0.94] | | 0.83 [0.78, 0.87] |
| 23 | Storandt & Morris (2010) | 55 | 23 | 34 | 9 | 21 | 14 | | 0.62 [0.49, 0.73] | | 0.61 [0.41, 0.78] |
| 24 | Thompson et al. (2011) | 27 | 156 | 22 | 16 | 5 | 140 | | 0.81 [0.63, 0.92] | | 0.90 [0.84, 0.94] |
| 25 | Vogel et al. (2007) * | 35 | 28 | 31 | 1 | 4 | 27 | | 0.89 [0.74, 0.95] | | 0.96 [0.82, 0.99] |
| 26 | Welsh et al. (1991) * | 147 | 49 | 98 | 2 | 49 | 47 | | 0.67 [0.59, 0.74] | | 0.96 [0.86, 0.99] |

* = *synthetic score*

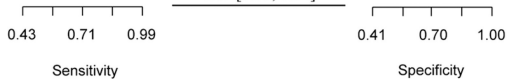


**Fig. 2.**

Paired forest plot AD vs HC Immediate Recall measures. AD: Alzheimer's disease, HC: healthy controls, TP: true positive, FP: false positive, FN: false negative, TN: true negative

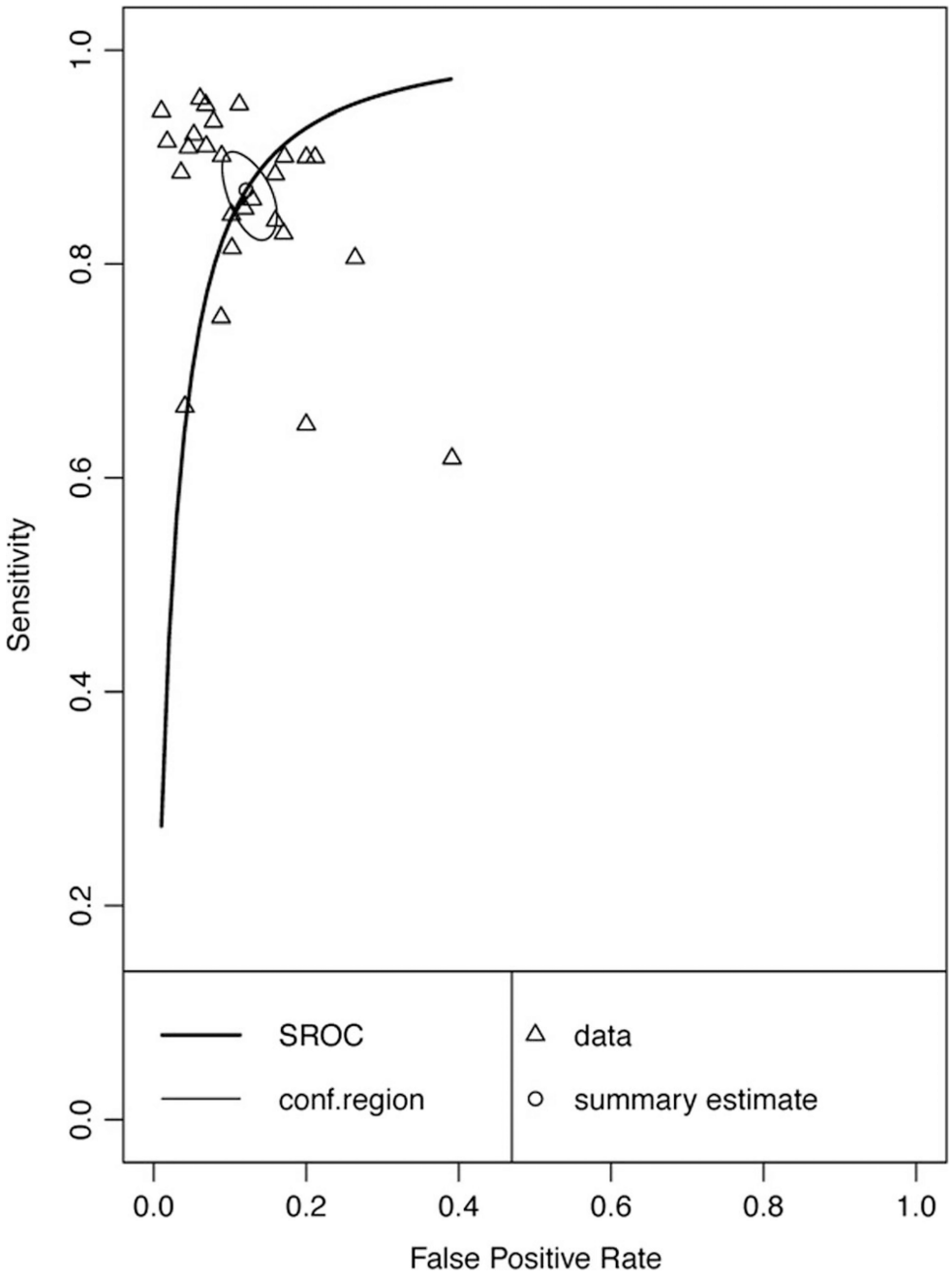


**Fig. 3.**

Hierarchical summary receiver-operator characteristic (SROC) curve for AD vs HC Immediate Recall measures. Conf.region = confidence region at the 95th percentile

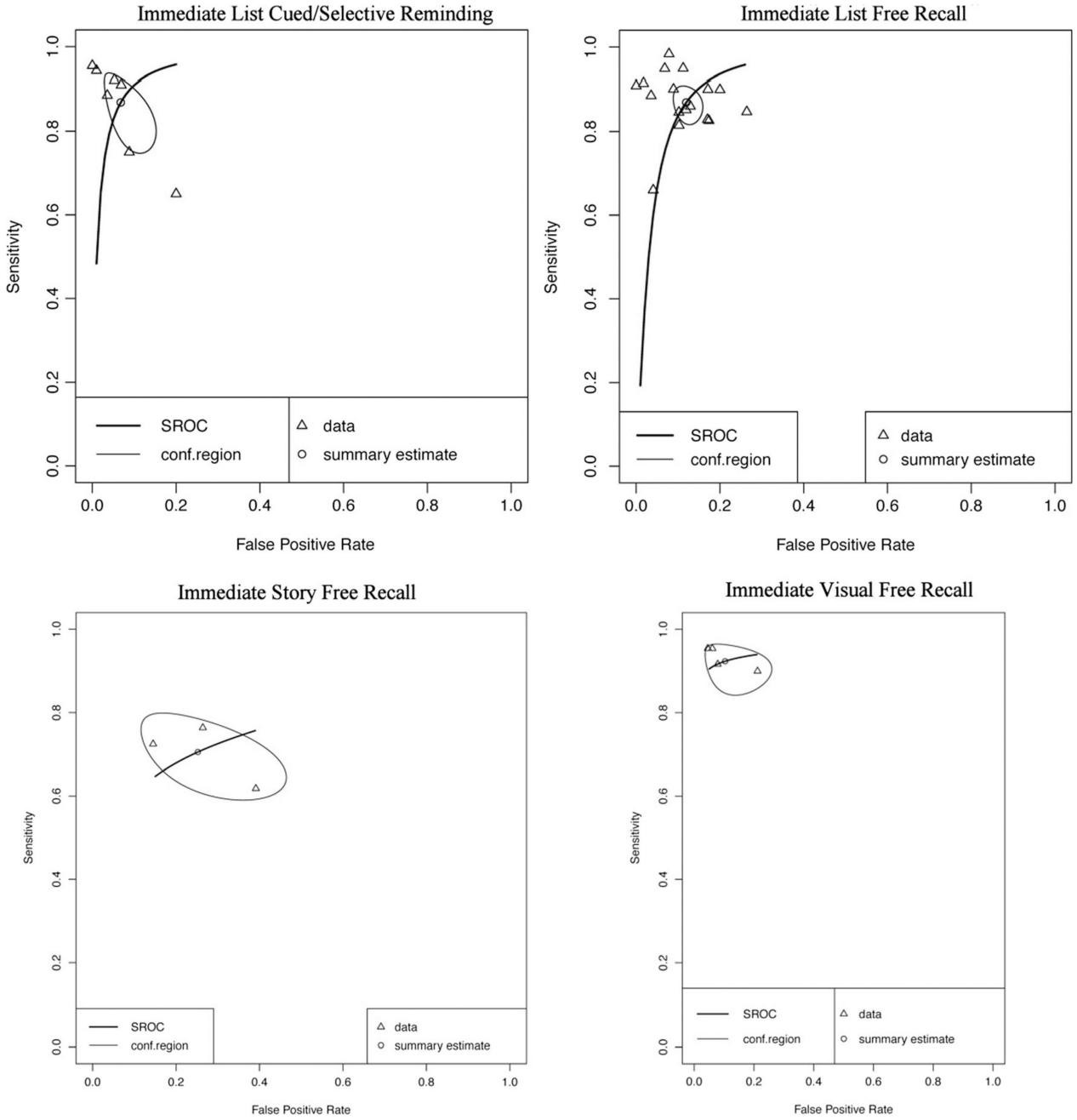


**Fig. 4.**

Hierarchical summary receiver-operator characteristic (SROC) curve for AD versus HC for subclasses of Immediate Recall measures. Conf.region = confidence region at the 95th percentile

| | Study | AD | HC | TP | FP | FN | TN | Sensitivity (95%, CI) | Sensitivity (95%, CI) | Specificity (95%, CI) | Specificity (95%, CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Baek et al. (2012) * | 72 | 53 | 62 | 10 | 10 | 43 | | 0.86 [0.76, 0.92] | | 0.81 [0.69, 0.89] |
| 2 | Bertolucci et al. (2001) | 43 | 85 | 32 | 15 | 11 | 70 | | 0.74 [0.60, 0.85] | | 0.82 [0.73, 0.89] |
| 3 | Cahn et al. (1995) | 42 | 236 | 37 | 38 | 5 | 198 | | 0.88 [0.75, 0.95] | | 0.84 [0.79, 0.88] |
| 4 | Chandler et al. (2005) | 95 | 95 | 88 | 9 | 7 | 86 | | 0.93 [0.86, 0.96] | | 0.91 [0.83, 0.95] |
| 5 | Clark et al. (2010) * | 73 | 20 | 68 | 1 | 5 | 19 | | 0.93 [0.85, 0.97] | | 0.95 [0.76, 0.99] |
| 6 | Delgado et al. (2016a) * | 50 | 42 | 47 | 3 | 3 | 39 | | 0.94 [0.84, 0.98] | | 0.93 [0.81, 0.98] |
| 7 | Delgado et al. (2016b) * | 50 | 42 | 45 | 3 | 5 | 39 | | 0.90 [0.79, 0.96] | | 0.93 [0.81, 0.98] |
| 8 | Duff et al. (2008) * | 69 | 69 | 63 | 6 | 6 | 63 | | 0.91 [0.82, 0.96] | | 0.91 [0.82, 0.96] |
| 9 | Ewers et al. (2012) | 81 | 101 | 70 | 10 | 11 | 91 | | 0.86 [0.77, 0.92] | | 0.90 [0.83, 0.95] |
| 10 | Gavett et al. (2009) | 26 | 98 | 24 | 3 | 2 | 95 | | 0.92 [0.76, 0.98] | | 0.97 [0.91, 0.99] |
| 11 | Gonzalez-Palau et al. (2013) | 54 | 109 | 48 | 21 | 6 | 88 | | 0.89 [0.78, 0.95] | | 0.81 [0.72, 0.87] |
| 12 | Ivanoiu et al. (2005) * | 22 | 22 | 21 | 3 | 1 | 19 | | 0.95 [0.78, 0.99] | | 0.86 [0.67, 0.95] |
| 13 | Lehrner et al. (2007) | 78 | 58 | 73 | 5 | 5 | 53 | | 0.94 [0.86, 0.97] | | 0.91 [0.81, 0.96] |
| 14 | Lemos et al. (2014) | 70 | 101 | 67 | 3 | 3 | 98 | | 0.96 [0.88, 0.99] | | 0.97 [0.92, 0.99] |
| 15 | Loewenstein et al. (2004) * | 26 | 53 | 21 | 6 | 5 | 47 | | 0.81 [0.62, 0.91] | | 0.89 [0.77, 0.95] |
| 16 | Park et al. (2016) * | 28 | 34 | 24 | 3 | 4 | 31 | | 0.86 [0.69, 0.94] | | 0.91 [0.77, 0.97] |
| 17 | Parra et al. (2010) | 22 | 30 | 17 | 1 | 5 | 29 | | 0.77 [0.57, 0.90] | | 0.97 [0.83, 0.99] |
| 18 | Paulsen et al. (1995) | 20 | 20 | 18 | 1 | 2 | 19 | | 0.90 [0.70, 0.97] | | 0.95 [0.76, 0.99] |
| 19 | Saka et al. (2006) * | 45 | 33 | 42 | 2 | 3 | 31 | | 0.93 [0.82, 0.98] | | 0.94 [0.80, 0.98] |
| 20 | Salmon et al. (2002) * | 98 | 98 | 90 | 13 | 8 | 85 | | 0.92 [0.85, 0.96] | | 0.87 [0.79, 0.92] |
| 21 | Shi et al. (2014) | 97 | 249 | 95 | 15 | 2 | 234 | | 0.98 [0.93, 0.99] | | 0.94 [0.90, 0.96] |
| 22 | Sotaniemi et al. (2012) * | 171 | 315 | 145 | 63 | 26 | 252 | | 0.85 [0.79, 0.89] | | 0.80 [0.75, 0.84] |
| 23 | Takechi & Dodge (2010) * | 128 | 54 | 108 | 7 | 20 | 47 | | 0.84 [0.77, 0.90] | | 0.87 [0.76, 0.94] |
| 24 | Testa et al. (2004) * | 306 | 409 | 245 | 70 | 61 | 339 | | 0.80 [0.75, 0.84] | | 0.83 [0.79, 0.86] |
| 25 | Thompson et al. (2011) | 27 | 156 | 26 | 16 | 1 | 140 | | 0.96 [0.82, 0.99] | | 0.90 [0.84, 0.94] |
| 26 | Vogel et al. (2007) * | 35 | 28 | 33 | 2 | 2 | 26 | | 0.94 [0.81, 0.98] | | 0.93 [0.77, 0.98] |
| 27 | Welsh et al. (1991) * | 147 | 49 | 123 | 2 | 24 | 47 | | 0.84 [0.77, 0.89] | | 0.96 [0.86, 0.99] |

* = *synthetic score*

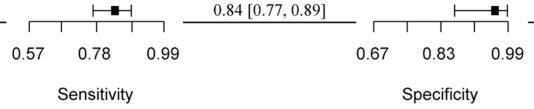


**Fig. 5.**

Paired forest plot AD vs HC Delayed Recall measures. AD: Alzheimer's disease, HC: healthy controls, TP: true positive, FP: false positive, FN: false negative, TN: true negative

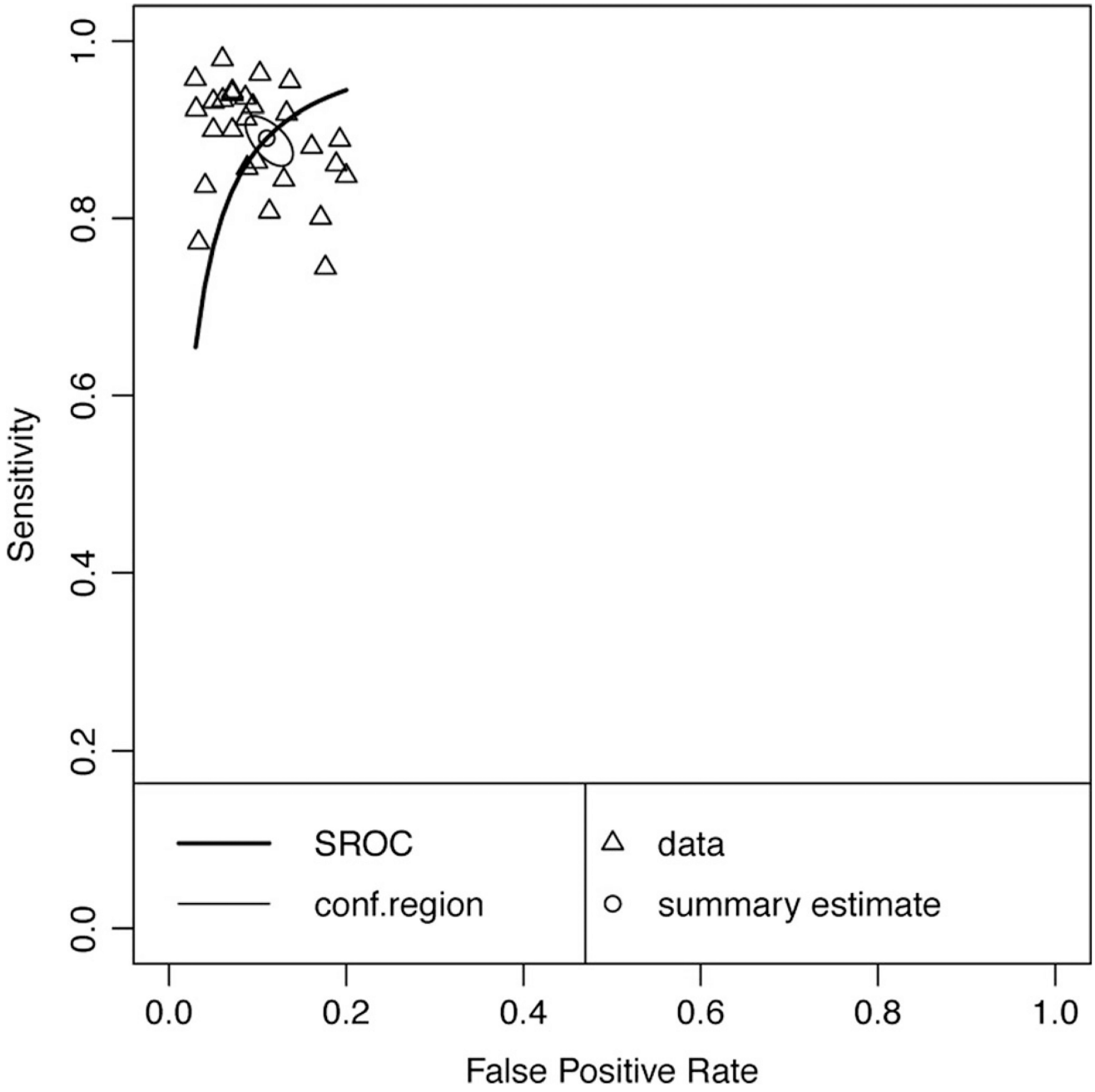


**Fig. 6.**

Hierarchical summary receiver-operator characteristic (SROC) curve for AD vs HC Delayed Recall measures. Conf.region = confidence region at the 95th percentile

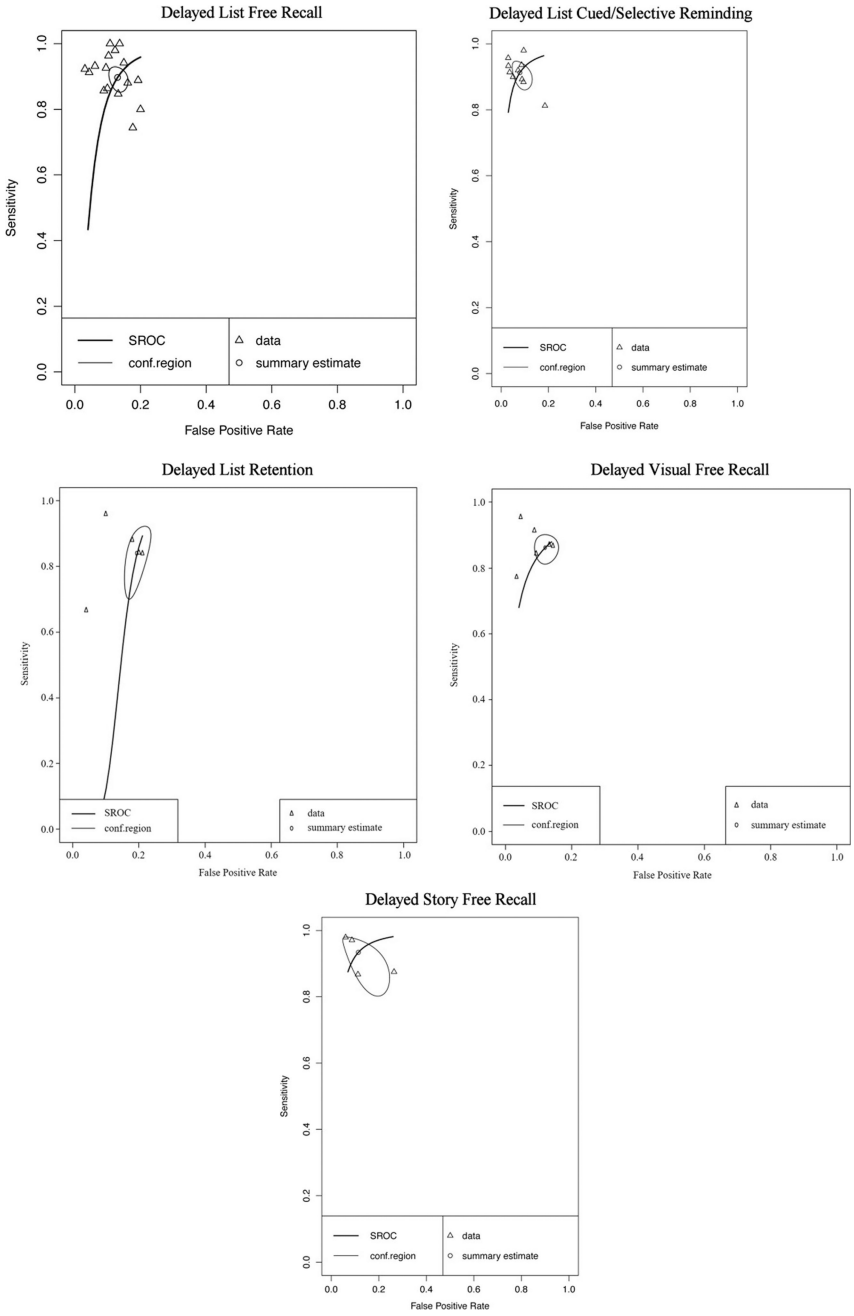


**Fig. 7.**

Hierarchical summary receiver-operator characteristic (SROC) curve for AD vs HC for subclasses of Delayed Recall measures. Conf.region = confidence region at the 95th percentile

| | Study | MCI | HC | TP | FP | FN | TN | Sensitivity (95%, CI) | Specificity (95%, CI) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Baek et al. (2011) | 112 | 53 | 76 | 24 | 36 | 29 | 0.68 [0.59, 0.76] | 0.55 [0.41, 0.67] |
| 2 | Blanco-Campal et al. (2009)* | 19 | 21 | 16 | 5 | 3 | 16 | 0.84 [0.62, 0.94] | 0.76 [0.55, 0.89] |
| 3 | Crocco et al. (2014) | 34 | 47 | 30 | 5 | 4 | 42 | 0.88 [0.73, 0.95] | 0.89 [0.77, 0.95] |
| 4 | De Jager et al. (2003) | 29 | 51 | 22 | 11 | 7 | 40 | 0.76 [0.58, 0.88] | 0.78 [0.65, 0.88] |
| 5 | Duff et al. (2010)* | 72 | 71 | 22 | 11 | 50 | 60 | 0.31 [0.21, 0.42] | 0.85 [0.74, 0.91] |
| 6 | Gavett et al. (2009)* | 29 | 98 | 21 | 33 | 8 | 65 | 0.72 [0.54, 0.85] | 0.66 [0.57, 0.75] |
| 7 | Gonzalez-Palau et al. (2013) | 132 | 109 | 92 | 28 | 40 | 81 | 0.70 [0.61, 0.77] | 0.74 [0.65, 0.82] |
| 8 | Lemos et al. (2015) | 100 | 101 | 72 | 17 | 28 | 84 | 0.72 [0.63, 0.80] | 0.83 [0.75, 0.89] |
| 9 | Loewenstein et al. (2006)* | 23 | 80 | 13 | 6 | 10 | 74 | 0.57 [0.37, 0.74] | 0.92 [0.85, 0.97] |
| 10 | Park et al. (2016) | 29 | 34 | 17 | 8 | 12 | 26 | 0.59 [0.41, 0.74] | 0.76 [0.60, 0.88] |
| 11 | Parra et al. (2012) | 10 | 10 | 8 | 2 | 2 | 8 | 0.80 [0.49, 0.94] | 0.80 [0.49, 0.94] |
| 12 | Pike et al. (2013)* | 77 | 77 | 66 | 6 | 11 | 71 | 0.86 [0.76, 0.92] | 0.92 [0.84, 0.96] |
| 13 | Rabin et al. (2009)* | 38 | 51 | 33 | 13 | 5 | 38 | 0.87 [0.73, 0.94] | 0.75 [0.61, 0.84] |
| 14 | Ritter et al. (2006) | 20 | 25 | 7 | 2 | 13 | 23 | 0.35 [0.18, 0.57] | 0.92 [0.75, 0.98] |
| 15 | Schrijnemaekers et al. (2006) | 19 | 54 | 16 | 11 | 3 | 43 | 0.84 [0.62, 0.94] | 0.80 [0.67, 0.88] |
| 16 | Shankle et al. (2005) | 95 | 119 | 81 | 11 | 14 | 108 | 0.85 [0.77, 0.91] | 0.91 [0.84, 0.95] |
| 17 | Shi et al. (2012) | 134 | 249 | 92 | 72 | 42 | 177 | 0.69 [0.60, 0.76] | 0.71 [0.65, 0.76] |

* = *synthetic score*

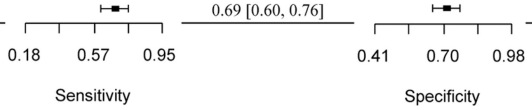


**Fig. 8.**

Paired forest plot MCI vs HC Immediate Recall measures. MCI: Mild Cognitive Impairment, HC: healthy controls, TP: true positive, FP: false positive, FN: false negative, TN: true negative. Conf.region = confidence region at the 95th percentile

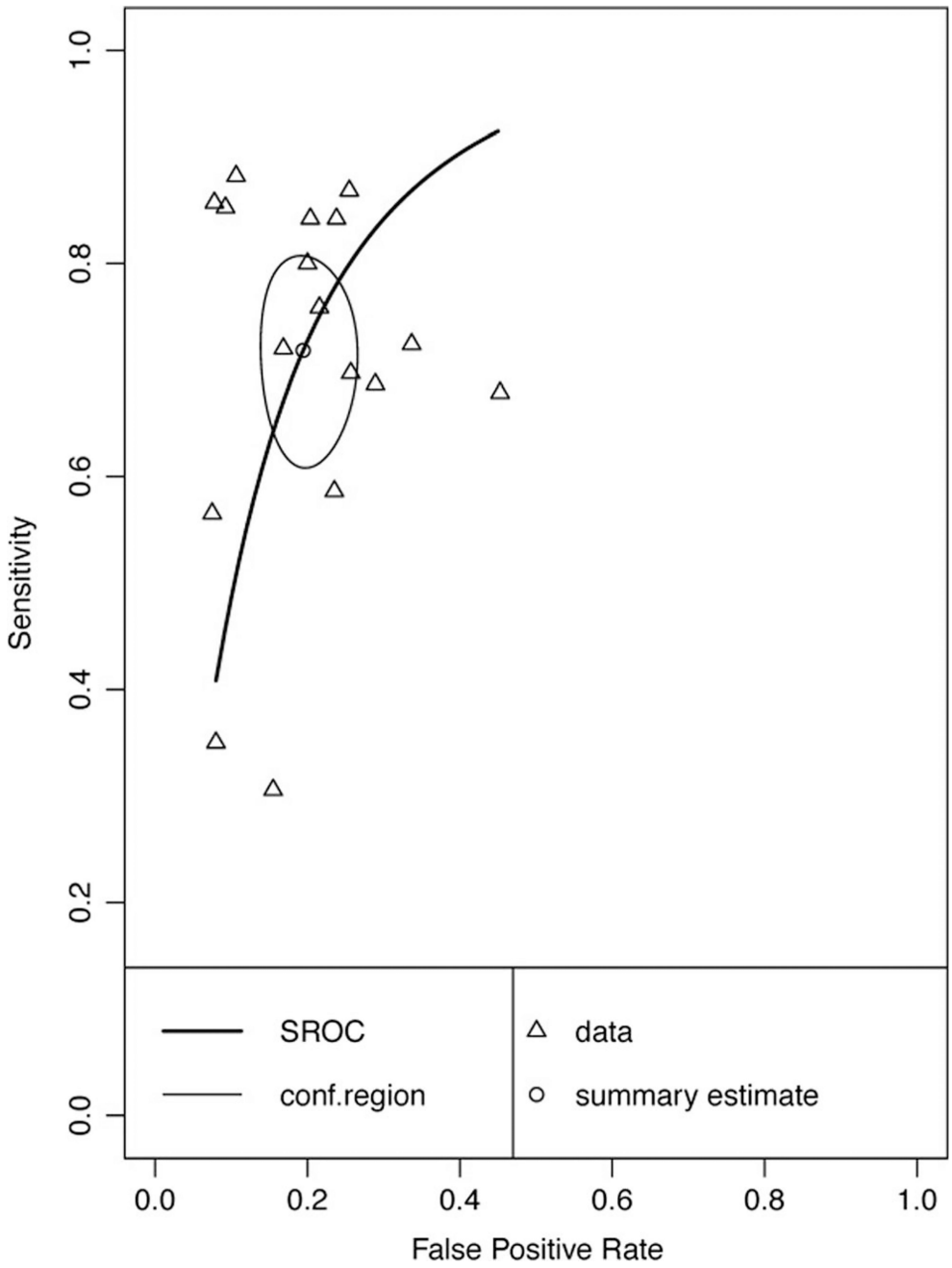


**Fig. 9.**

Hierarchical summary receiver-operator characteristic (SROC) curve for MCI vs HC Immediate Recall measures

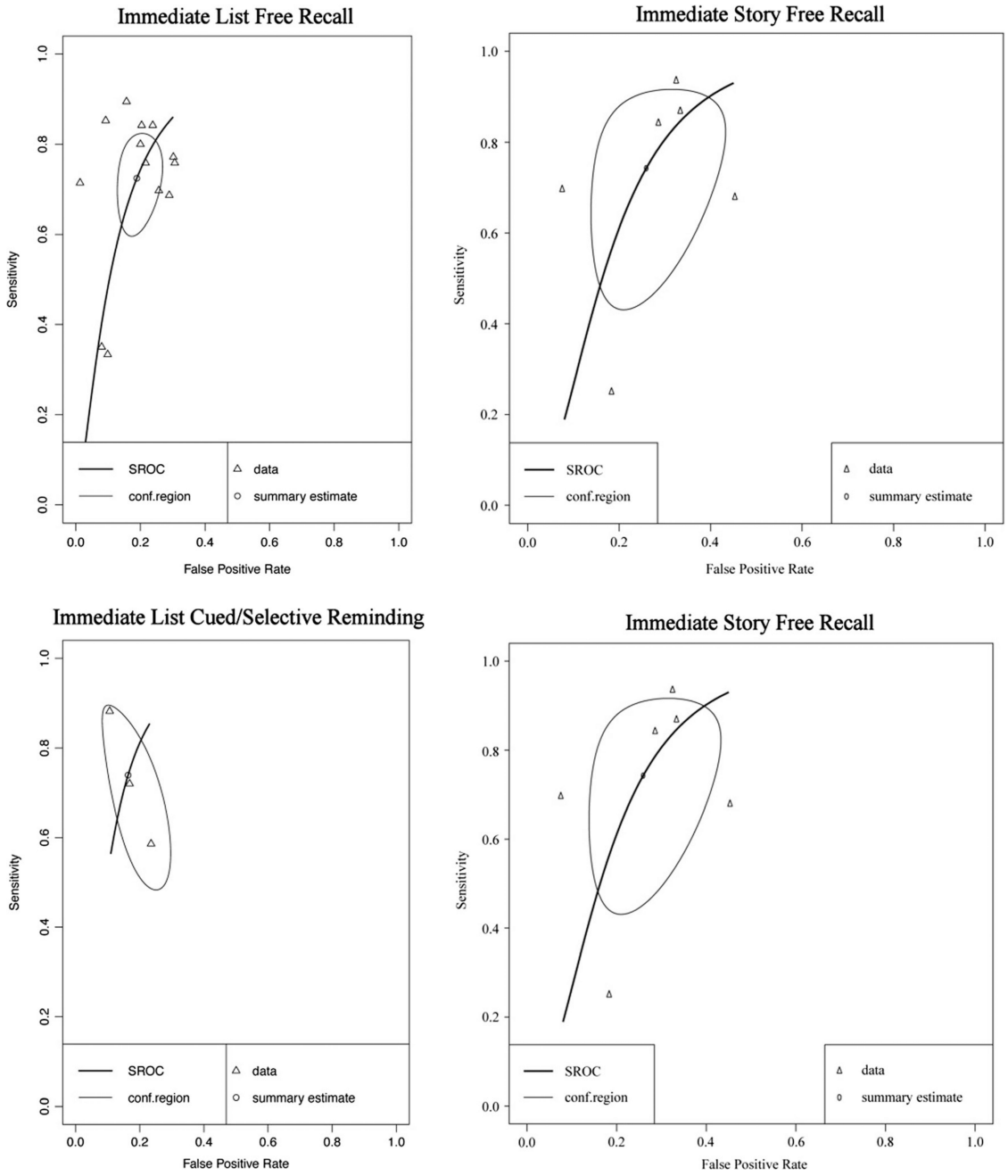


**Fig. 10.**

Hierarchical summary receiver-operator characteristic (SROC) curve for MCI vs HC for subclasses of Immediate Recall measures. Conf.region = confidence region at the 95th percentile

| | Study | MCI | HC | TP | FP | FN | TN | Sensitivity (95%, CI) | Sensitivity (95%, CI) | Specificity (95%, CI) | Specificity (95%, CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Baek et al. (2011) | 112 | 53 | 91 | 23 | 21 | 30 | | 0.81 [0.73, 0.87] | | 0.57 [0.43, 0.69] |
| 2 | Blanco-Campal et al. (2009)* | 19 | 21 | 13 | 3 | 6 | 18 | | 0.68 [0.46, 0.85] | | 0.86 [0.65, 0.95] |
| 3 | Chandler et al. (2005) | 60 | 95 | 49 | 35 | 11 | 60 | | 0.82 [0.70, 0.89] | | 0.63 [0.53, 0.72] |
| 4 | Clark et al. (2010)* | 44 | 20 | 34 | 3 | 10 | 17 | | 0.77 [0.63, 0.87] | | 0.85 [0.64, 0.95] |
| 5 | Delprado et al. (2012) | 84 | 84 | 74 | 13 | 10 | 71 | | 0.88 [0.79, 0.93] | | 0.85 [0.75, 0.91] |
| 6 | Duff et al. (2010)* | 72 | 71 | 29 | 10 | 43 | 61 | | 0.40 [0.30, 0.52] | | 0.86 [0.76, 0.92] |
| 7 | Gavett et al. (2009)* | 29 | 98 | 22 | 10 | 7 | 88 | | 0.76 [0.58, 0.88] | | 0.90 [0.82, 0.94] |
| 8 | Gonzalez-Palau et al. (2013) | 132 | 109 | 99 | 21 | 33 | 88 | | 0.75 [0.67, 0.82] | | 0.81 [0.72, 0.87] |
| 9 | Junkkila et al. (2012) | 17 | 22 | 11 | 3 | 6 | 19 | | 0.65 [0.41, 0.83] | | 0.86 [0.67, 0.95] |
| 10 | Karantzoulis et al. (2013) | 81 | 81 | 58 | 11 | 23 | 70 | | 0.72 [0.61, 0.80] | | 0.86 [0.77, 0.92] |
| 11 | Lekeu et al. (2010) | 34 | 14 | 30 | 4 | 4 | 10 | | 0.88 [0.73, 0.95] | | 0.71 [0.45, 0.88] |
| 12 | Lemos et al. (2015) | 100 | 101 | 76 | 19 | 24 | 82 | | 0.76 [0.67, 0.83] | | 0.81 [0.72, 0.88] |
| 13 | Loewenstein et al. (2004)* | 53 | 53 | 39 | 11 | 14 | 42 | | 0.74 [0.60, 0.84] | | 0.79 [0.67, 0.88] |
| 14 | Loewenstein et al. (2006)* | 23 | 80 | 16 | 7 | 7 | 73 | | 0.70 [0.49, 0.84] | | 0.91 [0.83, 0.96] |
| 15 | Park et al. (2016) * | 29 | 34 | 19 | 11 | 10 | 23 | | 0.66 [0.47, 0.80] | | 0.68 [0.51, 0.81] |
| 16 | Parra et al. (2012) | 10 | 10 | 8 | 2 | 2 | 8 | | 0.80 [0.49, 0.94] | | 0.80 [0.49, 0.94] |
| 17 | Pike et al. (2013)* | 77 | 77 | 69 | 15 | 8 | 62 | | 0.90 [0.81, 0.95] | | 0.81 [0.70, 0.88] |
| 18 | Rabin et al. (2009)* | 38 | 51 | 33 | 11 | 5 | 40 | | 0.87 [0.73, 0.94] | | 0.78 [0.65, 0.88] |
| 19 | Saka et al. (2006)* | 45 | 33 | 21 | 4 | 24 | 29 | | 0.47 [0.33, 0.61] | | 0.88 [0.73, 0.95] |
| 20 | Serna et al. (2015) | 352 | 890 | 201 | 258 | 151 | 632 | | 0.57 [0.52, 0.62] | | 0.71 [0.68, 0.74] |
| 21 | Shankle et al. (2005) | 95 | 119 | 78 | 11 | 17 | 108 | | 0.82 [0.73, 0.89] | | 0.91 [0.84, 0.95] |
| 22 | Shi et al. (2014) | 134 | 249 | 121 | 50 | 13 | 199 | | 0.90 [0.84, 0.94] | | 0.80 [0.75, 0.84] |

* = *synthetic score*


**Fig. 11.**

Paired forest plot MCI vs HC Delayed Recall measures. MCI: Mild Cognitive Impairment, HC: healthy controls, TP: true positive, FP: false positive, FN: false negative, TN: true negative

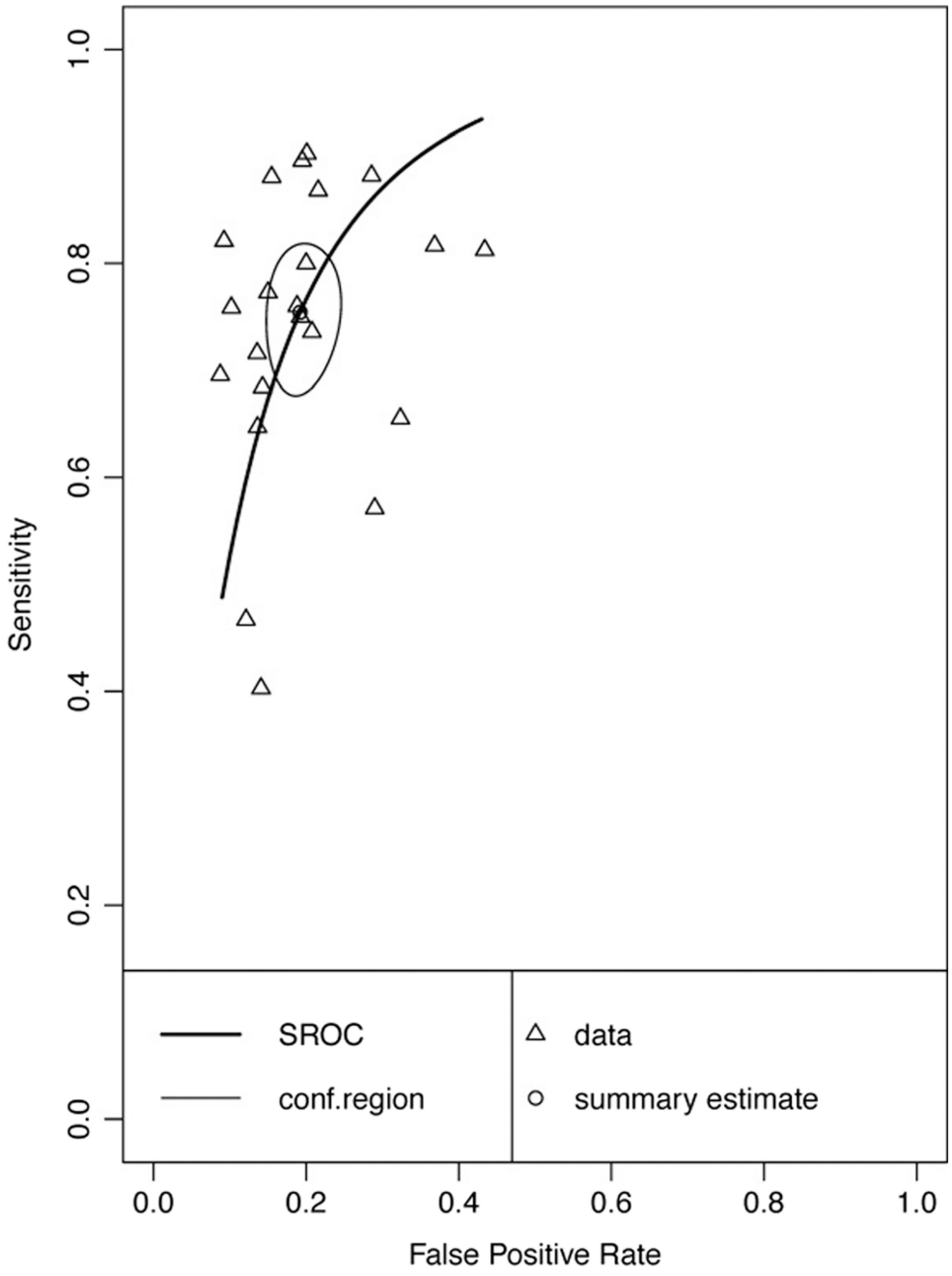


**Fig. 12.**

Hierarchical summary receiver-operator characteristic (SROC) curve for MCI vs HC Delayed Recall measures. Conf.region = confidence region at the 95th percentile

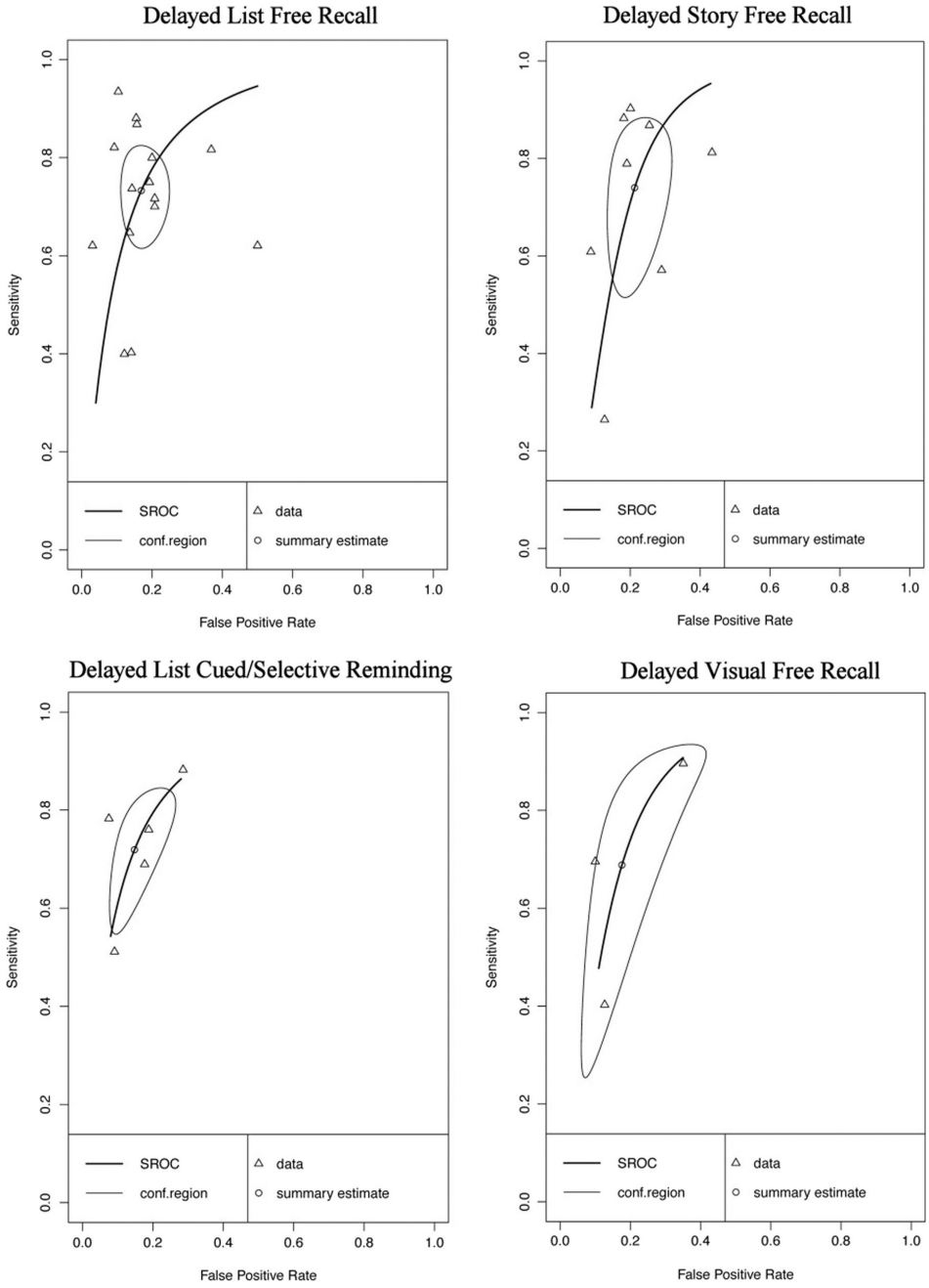


**Fig. 13.**

Hierarchical summary receiver-operator characteristic (SROC) curve for MCI vs HC for subclasses of Delayed Recall measures. Conf.region = confidence region at the 95th percentile

**Table 1**

Meta-analyses of immediate recall measures for Alzheimer's disease vs. healthy controls

| **Univariate Analysis** | **Immediate Recall Measures** | **Immediate List Free Recall** | **Immediate List Cued/Selective Reminding** | **Immediate Story Free Recall** | **Immediate Visual Free Recall** |
|---|---|---|---|---|---|
| k | 26 | 17 | 7 | 3 | 4 |
| Equality of sensitivities | $\chi^2(25) = 114.84$ $p. < .0001$ | $\chi^2(16) = 75.44$ $p. < .0001$ | $\chi^2(6) = 19.72$ $p. = .003$ | $\chi^2(2) = 3.33$ $p. = .189$ | $\chi^2(3) = .95$ $p. = .812$ |
| Equality of specificities | $\chi^2(25) = 104.10$ $p. < .0001$ | $\chi^2(16) = 53.30$ $p. < .0001$ | $\chi^2(6) = 14.56$ $p. = .024$ | $\chi^2(2) = 6.59$ $p. = .037$ | $\chi^2(3) = 11.33$ $p. = .010$ |
| *Rho* (Se and false positive rate correlation) (95% CI) | −.53 (−.76, −.18) | −.09 (−.55, .41) | −.93 (−.99, −.61) | NA | −.85 (−1.0, .61) |
| DOR (95% CI) | 56.18 (35.63, 88.58) | 55.98 (35.38, 88.58) | 95.31 (26.76, 339.45) | 7.29 (2.70, 19.71) | 105.45 (32.30, 344.25) |
| Cochran's Q | $Q(25) = 31.57$ $p. = .171$ | $Q(16) = 18.12$ $p. = .317$ | $Q(6) = 6.06$ $p. = .417$ | $Q(2) = 2.18$ $p. = .335$ | $Q(3) = 2.74$ $p. = .434$ |
| Tau (95% CI) | .97 (0.00, 1.46) | .74 (0.00, 1.31) | 1.46 (0.00, 3.42) | .75 (0.00, 5.80) | .81 (0.00, 4.43) |
| Tau-squared (95% CI) | .95 (0.00, 2.13) | .55 (0.00, 1.72) | 2.13 (0.00, 11.68) | .57 (0.00, 33.68) | .65 (0.00, 19.61) |
| Meta-analysis | | | | | |
| Sensitivity (95% CI) | .87 (.83, .90) | .87 (.83, .90) | .87 (.78, .93) | .71 (.61, .78) | .92 (.86, .96) |
| Specificity (95% CI) | .88 (.85, .90) | .88 (.85, .91) | .93 (.87, .97) | .75 (.58, .86) | .90 (.78, .95) |

## Table 2

Sensitivity analyses of Alzheimer's disease vs. healthy control studies where the measure of interest was not used in participant diagnosis

| Univariate Analysis | Immediate Recall Measures | Delayed Recall Measures |
|---|---|---|
| k | 18 | 19 |
| Equality of sensitivities | $\chi^2(17) = 56.30$ $p. < .0001$ | $\chi^2(18) = 47.67$ $p. < .001$ |
| Equality of specificities | $\chi^2(17) = 71.57$ $p. < .0001$ | $\chi^2(18) = 45.75$ $p. < .001$ |
| Rho (Se and false positive rate correlation) (95% CI) | −.71 (−.88, −.36) | −.35 (−.69, .13) |
| DOR (95% CI) | 56.33 (30.03, 105.66) | 75.36 (44.77, 126.85) |
| Cochran's Q | Q(17) = 16.74 $p. = .472$ | Q(18) = 15.53 $p. = .625$ |
| Tau (95% CI) | 1.16 (0.00, 1.49) | .91 (0.00, 1.06) |
| Tau-squared (95% CI) | 1.35 (0.00, 2.21) | .82 (0.00, 1.12) |
| Meta-analysis | | |
| Sensitivity (95% CI) | .86 (.82, .90) | .89 (.85, .91) |
| Specificity (95% CI) | .88 (.84, .92) | .89 (.86, .91) |

**Table 3**

Meta-analyses of delayed recall measures for Alzheimer's disease vs. healthy controls

| Univariate Analysis | Delayed Recall Measures | Delayed List Free Recall | Delayed List Cued/ Selective Reminding | Delayed List Retention | Delayed Visual Free Recall | Delayed Story Free Recall |
|---|---|---|---|---|---|---|
| k | 27 | 16 | 10 | 5 | 6 | 4 |
| Equality of sensitivities | $\chi^2(26) = 63.40$ $p. < .0001$ | $\chi^2(15) = 38.54$ $p. < .001$ | $\chi^2(9) = 19.62$ $p. = .020$ | $\chi^2(4) = 35.40$ $p. < .0001$ | $\chi^2(5) = 5.06$ $p. = .409$ | $\chi^2(3) = 13.12$ $p. = .004$ |
| Equality of specificities | $\chi^2(26) = 77.89$ $p. < .0001$ | $\chi^2(15) = 26.71$ $p. = .031$ | $\chi^2(9) = 14.46$ $p. = .107$ | $\chi^2(4) = 9.61$ $p. = .048$ | $\chi^2(5) = 5.25$ $p. = .386$ | $\chi^2(3) = 20.87$ $p. < .001$ |
| *Rho* (Se and false positive rate correlation) (95% CI) | −.38 (−.66, .00) | −.44 (−.77, .07) | −.68 (−.92, −.08) | .46 (−.71, .95) | .11 (−.77, .85) | −.69 (−.99, .80) |
| DOR (95% CI) | 78.41 (51.32, 119.80) | 69.35 (42.33, 113.60) | 146.01 (60.21, 354.08) | 27.40 (16.90, 44.42) | 56.54 (35.21, 90.79) | 113.47 (24.42, 527.14) |
| Cochran's Q | $Q(26) = 24.26$ $p. = .561$ | $Q(15) = 15.75$, $p. = .399$ | $Q(9) = 5.84$ $p. = .756$ | $Q(4) = 5.57$ $p. = .233$ | $Q(5) = 4.47$ $p. = .484$ | Q(3) = 3.44 $p. = .329$ |
| Tau (95% CI) | .88 (0.00, .95) | .75 (0.00, 1.32) | 1.12 (0.00, 1.43) | .35 (0.00, 2.94) | .00 (0.00, 2.11) | 1.43 (0.00, 6.08) |
| Tau-squared (95% CI) | .77 (0.00, .91) | .56 (0.00, 1.74) | 1.25 (0.00, 2.05) | .13 (0.00, 8.65) | .00 (0.00, 4.46) | 2.05 (0.00, 36.99) |
| Meta-analysis | | | | | | |
| Sensitivity (95% CI) | .89 (.87, .91) | .90 (.86, .92) | .91 (.87, .94) | .84 (.73, .91) | .86 (.82, .89) | .93 (.84, .98) |
| Specificity (95% CI) | .89 (.87, .91) | .87 (.84, .89) | .92 (.88, .95) | .81 (.77, .84) | .88 (.85, .91) | .89 (.79, .94) |

**Table 4**

Meta-analyses of immediate recall measures for Mild Cognitive Impairment vs. healthy controls

| **Univariate Analysis** | **Immediate Recall Measures** | **Immediate List Free Recall** | **Immediate Story Free Recall** | **Immediate List Cued/Selective Reminding** |
|---|---|---|---|---|
| k | 17 | 13 | 6 | 3 |
| Equality of sensitivities | $\chi^2(16) = 104.82$ $p. < .0001$ | $\chi^2(12) = 84.60$ $p. < .0001$ | $\chi^2(5) = 92.32$ $p. < .0001$ | $\chi^2(2) = 7.10$ $p. = .029$ |
| Equality of specificities | $\chi^2(16) = 74.25$ $p. < .0001$ | $\chi^2(12) = 56.41$ $p. < .0001$ | $\chi^2(5 = 29.98$ $p. < .0001$ | $\chi^2(2) = 2.40$ $p. = .301$ |
| *Rho* (Se and false positive rate correlation) (95% CI) | .11 (−.39, .56) | .37 (−.23, .76) | .38 (−.63, .91) | NA |
| DOR (95% CI) | 11.19 (6.76, 18.53) | 12.76 (7.53, 21.64) | 8.55 (2.86, 25.54) | 14.26 (4.22, 48.26) |
| Cochran's Q | $Q(16) = 15.99$ $p. = .453$ | $Q(12) = 13.34$ $p. = .345$ | $Q(5) = 4.40$ $p. = .494$ | $Q(2) = 2.70$ $p. = .259$ |
| Tau (95% CI) | .90 (0.00, 1.19) | .77 (0.00, 1.47) | 1.25 (0.00, 2.78) | .93 (0.00, 8.23) |
| Tau-squared (95% CI) | .82 (0.00, 1.41) | .59 (0.00, 2.16) | 1.57 (0.00, 7.71) | .87 (0.00, 67.79) |
| Meta-analysis | | | | |
| Sensitivity (95% CI) | .72 (.63, .79) | .72 (.62, .81) | .74 (.50, .89) | .74 (.54, .87) |
| Specificity (95% CI) | .81 (.75, .85) | .81 (.75, .86) | .74 (.60, .84) | .84 (.73, .90) |

## Table 5

Sensitivity analyses of Mild Cognitive Impairment vs. healthy control studies where the measure of interest was not used in participant diagnosis

| Univariate Analysis | Immediate Recall Measures | Delayed Recall Measures |
|---|---|---|
| k | 13 | 16 |
| Equality of sensitivities | $\chi^2(12) = 87.65$ $p. < .0001$ | $\chi^2(15) = 133.69$ $p. < .0001$ |
| Equality of specificities | $\chi^2(12) = 48.96$ $p. < .0001$ | $\chi^2(15) = 71.99$ $p. < .0001$ |
| Rho (Se and false positive rate correlation) (95% CI) | .08 (−.49, .60) | .04 (−.47, .52) |
| DOR (95% CI) | 11.69 (6.60, 20.70) | 14.53 (8.03, 26.29) |
| Cochran's Q | $Q(12) = 11.34$ $p. = .50$ | $Q(15) = 8.16$ $p. = .917$ |
| Tau (95% CI) | .89 (0.00, 1.27) | 1.09 (0.00, .64) |
| Tau-squared (95% CI) | .80 (0.00, 1.61) | 1.18 (0.00, .41) |
| Meta-analysis | | |
| Sensitivity (95% CI) | .73 (.63, .81) | .76 (.68, .82) |
| Specificity (95% CI) | .80 (.75, .85) | .81 (.77, .85) |

**Table 6**

Meta-analyses of delayed recall measures for Mild Cognitive Impairment vs. healthy controls

| **Univariate Analysis** | **Delayed Recall Measures** | **Delayed List Free Recall** | **Delayed Story Free Recall** | **Delayed List Cued/Selective Reminding** | **Delayed Visual Recall** |
|---|---|---|---|---|---|
| k | 22 | 15 | 8 | 5 | 3 |
| Equality of sensitivities | $\chi^2(21) = 160.59$ *p.* < .0001 | $\chi^2(14) = 105.45$ *p.* < .0001 | $\chi^2(7) = 138.44$ *p.* < .0001 | $\chi^2(4) = 15.65$ *p.* = .004 | $\chi^2(2) = 40.64$ *p.* < .0001 |
| Equality of specificities | $\chi^2(21) = 93.22$ *p.* < .0001 | $\chi^2(14) = 75.94$ *p.* < .0001 | $\chi^2(7) = 39.23$ *p.* < .0001 | $\chi^2(4) = 7.96$ *p.* = .093 | $\chi^2(2) = 18.73$ *p.* = < .0001 |
| *Rho* (Se and false positive rate correlation) (95% CI) | .22 (−.22, .59) | .02 (−.50, .52) | .36 (−.46, .85) | .64 (−.56, .97) | NA |
| DOR (95% CI) | 13.61 (8.63, 21.45) | 14.31 (8.14, 25.16) | 11.00 (4.62, 26.19) | 15.23 (9.49, 24.43) | 11.01 (4.35, 27.87) |
| Cochran's Q | *Q*(21) = 11.71 *p.* = .947 | *Q*(14) = 14.71 *p.* = .398 | *Q*(7) = 5.03 *p.* = .656 | *Q*(4) = 3.51 *p.* = .476 | Q(2) = 1.91 *p.* = .385 |
| Tau (95% CI) | .95 (0.00, 0.39) | .96 (0.00, 1.40) | 1.16 (0.00, 1.72) | 0.00 (0.00, 1.63) | .66 (0.00, 4.95) |
| Tau-squared (95% CI) | .91 (0.00, .15) | .92 (0.00, 1.95) | 1.34 (0.00, 2.95) | 0.00 (0.00, 2.64) | .44 (0.00, 24.48) |
| Meta-analysis | | | | | |
| Sensitivity (95% CI) | .75 (.69, .81) | .73 (.64, .81) | .74 (.56, .86) | .72 (.58, .82) | .69 (.33. .91) |
| Specificity) (95% CI) | .81 (.77, .84) | .83 (.77, .88) | .79 (.70, .85) | .85 (.76, .91) | .82 (.64, .92) |