ARTIFICIAL INTELLIGENCE AND NEUROPSYCHOLOGICAL MEASURES: THE CASE OF ALZHEIMER'S DISEASE

Petronilla Battista [a1], Christian Salvatore [b,c 1], Manuela Berlingeri [d], Antonio Cerasa [e,f], and Isabella Castiglioni [g,h*]

[a] Istituti Clinici Scientifici Maugeri, IRCCS, Pavia, Italy;

[b] Scuola Universitaria Superiore IUSS, Pavia, Italy.

[c] DeepTrace Technologies SRL, Milan, Italy

[d] Dipartimento di Studi Umanistici (DISTUM), Università degli Studi di Urbino Carlo Bo, Urbino (PU), Italy.

[e] IRIB, National Research Council, Mangone, CS, Italy

[f] S. Anna Institute and Research in Advanced Neurorehabilitation (RAN), Crotone, Italy

[g] Dipartimento di Fisica, Università degli Studi di Milano Bicocca, Milan, Italy

[h] Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Segrate, Milan, Italy

[1] equally contributed

***Corresponding author*:** Isabella Castiglioni
Institute of Molecular Bioimaging and Physiology,
National Research Council (IBFM-CNR),
Via F.lli Cervi, 93, 20090 Segrate, Milano, Italy
Tel. 02 21717511 Fax 02 21717558; email: isabella.castiglioni@ibfm.cnr.it

Co-Author emails:

petronillabattista@gmail.com

christian.salvatore@iusspavia.it

manuela.berlingeri@uniurb.it

a.cerasa@unicz.it

ORCID:
Petronilla Battista: 0000-0002-3120-1214
Christian Salvatore: 0000-0001-9312-4675
Manuela Berlingeri: 0000-0002-3159-2809
Antonio Cerasa: 0000-0002-8022-4770
Isabella Castiglioni: 0000-0001-7191-5417

If you wish to cite this article, please use the following reference:
Battista, P., Salvatore, C., Berlingeri, M., Cerasa, A., & Castiglioni, I. (2020). Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. Neuroscience and Biobehavioral Reviews, 114, 211–228. https://doi.org/10.1016/j.neubiorev.2020.04.026

ARTIFICIAL INTELLIGENCE AND NEUROPSYCHOLOGICAL MEASURES: THE CASE OF ALZHEIMER'S DISEASE

ABSTRACT

One of the current challenges in the field of Alzheimer's disease (AD) is to identify patients with mild cognitive impairment (MCI) that will convert to AD. Artificial intelligence, in particular machine learning (ML), has established as one of more powerful approach to extract reliable predictors and to automatically classify different AD phenotypes. It is time to accelerate the translation of this knowledge in clinical practice, mainly by using low-cost features originating from the neuropsychological assessment.

We performed a meta-analysis to assess the contribution of ML and neuropsychological measures for the automated classification of MCI patients and the prediction of their conversion to AD. The pooled sensitivity and specificity of patients' classifications was obtained by means of a quantitative bivariate random-effect meta-analytic approach. Although a high heterogeneity was observed, the results of meta-analysis show that ML applied to neuropsychological measures can lead to a successful automatic classification, being more specific as screening rather than prognosis tool. Relevant categories of neuropsychological tests can be extracted by ML that maximize the classification accuracy.

Keywords: Mild Cognitive impairment; MCI; AD; neurodegenerative diseases: dementia; biomarkers; neuropsychological tests; cognitive measures; machine learning; automatic classification.

1. INTRODUCTION

It was estimated that in 2015 there would have been more than 47 million people worldwide affected by dementia; these estimates were confirmed and the projections for 2050 are even more worrying with 131 million people living with dementia (Prince et al., 2013). This high prevalence has led to significant health and social problems and is expected to rise due to the increase in life expectancy and under- or mis-diagnosis.

Several forms of dementia have been described in the literature with Alzheimer's disease (AD) being considered the primary cause of neurodegenerative dementia (Querfurth and LaFerla 2010). Pathologically, this neurodegenerative disease has been linked by protein misfolding in the brain, with specific abnormal protein and pattern of deposition, which can occur years or even decades before clinical manifestation. However, currently, the only definitive diagnosis can be performed in *post-mortem* examination by detecting the presence of senile plaques and neurofibrillary tangles associated with amyloid angiopathy in the brain tissues (Beach et al., 2012).

In living human brain, the criteria for the diagnosis of AD are those proposed by the National Institute of Neurological Disorders and Stroke–Alzheimer Diseases and Related Disorders working group (McKhann et al., 1984). Since their publication, thanks to biological advances and neuroimaging studies on AD, several other international criteria have been proposed (Albert et al., 2011; Dubois et al., 2007, 2014; McKhann et al., 2011; Sperling et al., 2011). The National Institute on Aging and the Alzheimer's Association define different stages of AD progression (Jack et al., 2011), starting from the asymptomatic pre-clinical and the symptomatic prodromal stages. The stage of AD identified by merely subjective cognitive impairment (SCI) which cannot be detected by objective measures (Lautenschlager et al., 2005) can emerge several years before the onset of the prodromal stage (Solfrizzi et al., 2004). The subsequent stage (prodromal) displaying clear symptoms is referred to as Mild Cognitive

Impairment (MCI), which is characterized by symptoms that are not severe enough to meet currently accepted diagnostic criteria for AD (Petersen et al., 2004). Indeed, the term MCI is applied to subjects with a deficit of at least one cognitive domain, without significant effects on their daily activities (Albert et al., 2011). Finally, when cognitive and behavioral symptoms interfere with daily functional abilities, a diagnosis of AD can be made, with a label of probable or possible according to various clinical conditions (McKhann et al., 2011).

However, at the state of the art it is still difficult to predict patients at risks of AD (MCI) and whether and when individuals at risk (with MCI) will progress to AD-type dementia and how much time will lapse for progression. Thus, the current challenge is to identify markers that capture MCI and discriminate between patients with MCI who will convert (MCI converters, MCIc) and who will not convert (MCI not-converters, MCInc) to AD-type dementia.

In the last ten years, the international neuroimaging community has made considerable efforts to identify surrogate biomarkers of AD pathophysiology to be used for early (pre-clinical/prodromal stages) diagnosis. According to Alzheimer's disease Neuroimaging Initiative (Mueller et al., 2005), an ideal AD biomarker should be simple to perform, reliable, minimally invasive/expensive and able to detect features of the pathophysiologic processes active in AD before symptom onset. The vast majority of these biomarkers have been carried out in neurobiology realm by means of very sophisticated technologies. Tau- or amyloid-aggregation within the brain, cortical hypo-metabolism, and hippocampal atrophy can be obtained, *in vivo,* by Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) (Frisoni et al., 2010; Bateman et al., 2012; Jack et al., 2013). However, the availability of these technologies only in some expert centres and financial constraints limit their adoption into routine clinical use. More importantly, despite they are reliable biomarkers of disease, they have relatively limited ability in the prediction of AD at the individual level (Arbabshirani et al., 2017; Petersen et al., 2001).

In the last few years, artificial intelligence has proven to be a new effective way in designing prognostic/diagnostic tools for improving the clinical practice of AD. In particular, machine learning (ML) methods, which are intelligent systems capable of *learning* complex relationships or patterns from empirical data and extracting predictive data models, have found fertile ground in the study of AD with promising results for biomarkers also at an early-stage (Bishop 2006; Orrù et al., 2012, Salvatore et al., 2016). ML methods have been applied on several features with ability in the prediction of which subjects with MCI will progress to AD. These include biological, neuroimaging data and neuropsychological testing. Neuroimaging is the most important realm where ML methods have widely been applied (Weiner et al., 2017). By combining, in a multivariate way, information hidden in the brain images of patients and invisible to a naked eye, ML methods can automatically classify an individual subject on the basis of image differences or similarities with images of known classes of subjects (Noirhomme et al., 2014; Bryan 2016). From the first seminal paper by Klöppel et al., (2008), in the last 10 years, a plethora of ML studies on neuroimaging have been published with the aim to reach the best accuracy level in the automated clinical diagnosis of AD. Whilst classification accuracy in discriminating AD from healthy controls (HC) range between 80 to 95% (Mateos-Pérez et al., 2018) the real challenge is to automatically classify between prodromal forms of disease. A recent review of more than 30 papers showed that the ML automatic classification of MCIc vs MCInc, based on neuroimaging data, can achieve a median accuracy, specificity and sensitivity of 70%, 66% and 75%, respectively (Salvatore et al., 2016).

ML has also been applied to biological data of AD and MCI patients. By combining different biological markers in a multivariate way, this approach has been used to identify a biological signature of AD. For example, it was shown that the cerebrospinal fluid calbindin combined with cerebrospinal fluid Aβ42 can automatically discriminate between mildly and very mildly

demented subjects from HC with a sensitivity and specificity > 80% (Craig-Schapiro et al., 2011). However, in the classification of MCIc vs MCInc, this combination of measures led to a good sensitivity (80%) but a very low specificity (44%), with a balanced accuracy of 62% (Yang et al., 2012).

At a cognitive and behavioural level, international Working Group guidelines have proposed a list of neuropsychological tests for the diagnosis of AD (Dubois et al., 2014). However, there is still no clear consensus on the specific composite measures to be used for the early diagnosis of AD, since the discussion is still open as to what the most sensitive/specific tests are for early-stage AD. Some authors have argued that stringent cut-off should be fixed in order to identify whether performance is impaired for MCI subjects (Gainotti et al., 2014). To date, studies on the application of ML to neuropsychological tests are increasingly emerging, since some evidence show that ML systems can support the clinical classification of AD patients when trained on neuropsychological measures (Weiner et al., 2017).

In such a changing and stimulating scenario, we aim at performing, for the first time, a meta-analytic evaluation of the contribution of ML and neuropsychological measures for the automated classification of AD and MCI patients and the prediction of MCIs' conversion to AD-type dementia. Specifically, our purpose is to establish, from the results reported by independent studies, whether ML algorithms, trained on a set of neuropsychological measures, could be used for the automatic diagnosis of MCI and to automatically predict conversion to AD-type dementia. We also discuss the advantages associated with the adoption of intelligent tools in the field of neuropsychological assessments for clinical and experimental neuropsychologists by underlying a number of methodological issues that should be taken under consideration for translating these powerful approaches into reliable clinical studies.

2. MATERIAL AND METHODS

2.1 Search strategy and selection criteria

This systematic review was conducted on papers published on the use of ML applied to neuropsychological assessment for the automatic classification of AD, MCI and prediction of conversion of MCI to Alzheimer's type dementia. The protocol was not registered, but it was structured in accordance with the PRISMA statement (Moher et al., 2009), so that the PICOS approach was used to identify the studies to be included in the review and meta-analyses. Criteria for including or excluding papers were determined a priori. Papers were considered for inclusion only if: (a) they were written in full-text English language in a peer-reviewed journal, (b) they were published from 2010 to the end of search July 15, 2018, (c) they included subjects with a primary diagnosis of AD according to McKhann's criteria (McKhann et al., 1984), or subsequent versions, or subjects with a primary diagnosis of MCI according to Petersen's criteria (Petersen et al., 1999) and subsequent modifications (Petersen 2004; Petersen and Negash 2008), or who had available the global score of the Clinical Dementia Rating (CDR, total score 0.5); and (d) they included at least the neuropsychological measures for the classification. Articles were excluded if: (a) they did not include neuropsychological measures in the classification process, (b) they did not perform a classification of subjects, (c) they did not provide any classification performance, and (d) they did consider subjects with a history of other neurological or psychiatric disorders such as Parkinson's disease, and stroke. The two authors screened the publications on their relevance for the review. The final resulting papers were considered eligible for the review. Major details about information sources, search strategies and study selection process can be found in Supplementary Material.

2.2 Data extraction strategy

The data collected from each article were categorized as: information on the first author and year of publication, the size of cohorts, the modalities of measures used for the classification, the classification algorithm, the method used to validate the classification, and the classification performance in terms of study-specific accuracy, study-specific specificity, study-specific sensitivity, and study-specific AUC. The final papers were set into four categories, according to the following groups of subjects used for the automatic classification: 1) MCI vs HC, 2) MCIc vs MCInc, 3) AD vs HC, and 4) other comparisons (comparisons not already included in the first three categories).

Neuropsychological tests included in the automatic classification. For all comparisons, we assessed the neuropsychological tests whose scores were most frequently used as input for the classification and whose overall accuracy and/or AUC was higher than 0.7, since, according to the literature (Belleville et al., 2017) an accuracy score higher than 0.7 is considered to be good. The resulting subset of neuropsychological tests was referred to as *optimal predictors*.

2.3 Risk of bias in Individual Studies

Following the Cochrane guidelines, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool (Whiting et al., 2011) was used to assess the methodological quality and the risk of bias of each study. This quality assessment allowed classifying studies as a having low, high or unknown risk of bias. We used a high-quality report subgroup for meta-analyses.

2.4 Meta-Analysis

The data stored in the "predetermined grid" described at point 1.4 (namely sample sizes, study-specific sensitivity, study-specific specificity and study-specific accuracy) were used to

compute, for each single study, the number of true positive, true negative, false positives and false negatives cases. These were used as raw data for the meta-analysis.

The statistical analyses were carried out in R-Studio (version 13.1) using the MADA package (Doebler and Holling 2015). In particular, we first explored the neuropsychological data by producing forest plots. By using the "madad" function we obtained univariate measures of heterogeneity for the meta-analytic sensitivity and specificity ($Sensitivity_m$, $Specificity_m$) and we computed the heterogeneity I^2 index as follows:

$$I^2 = [(\chi^2 - df)/ \chi^2]*100.$$

Furthermore, by adopting the "reitsma" function, we computed a bivariate empty model (i.e. with the intercept only) to obtain the meta-analytic AUC (AUC_m) parameter for each summary ROC curve (with 95% contour ellipsoid). Here we report an example of the syntax of the bivariate reitsma model:

AUC_{NPS_AD} =AUC(reitsma(NPS_AD, formula = cbind(tsens, tfpr) ~ 1))

Finally, to explore the effect of covariates, that could explain the level of between-studies heterogeneity, we run a bivariate random effect model for logit-transformed pairs of sensitivity and specificity. In this latter case, we entered as predictor the factor "comparison" (namely, "AD vs HC", "MCI vs HC" and "MCIc VS MCInc) in the "reitsma" function.

3. RESULTS

3.1 Study selection

The literature search yielded 203 papers related to the established timeframe (January 1, 2010 - July 15, 2018) from electronic databases. Two authors included further six papers from the

reference list of previously retrieved articles. A total of 209 papers were identified. Based on the abstract and title, the two authors selected potentially relevant articles. Six hits were excluded because they were not referenced journal articles, 19 were excluded because they were reviews, books and book chapters not reporting original results, which left 184 records. Another 46 papers were further excluded because: a) they focused on a different topic; b) automatic classification was not performed; c) different populations were investigated. At this point in the screening, there were 138 papers left. These papers were checked by studying the full-text to exclude papers that did not meet inclusion criteria when this was not directly apparent from the title and the abstract. At this step, 6 papers were excluded due to the lack of information about sensitivity, specificity and accuracy; 46 papers were excluded because they did not perform a classification; 18 papers were excluded because they did not include neuropsychological tests, 9 papers were excluded because a comparison was made with other neurodegenerative disorders (e.g. Parkinson's disease). Finally, 59 papers were selected as eligible, which were therefore included in this review (see Figure 1).

3.2 Study Characteristics

Grouping papers according to the classes used in the automatic classification generated 4 tables. Table 1 shows papers on the classification of MCI vs HC, Table 2 on MCIc vs MCInc, Table 3 on AD vs HC, (for a total of 45 papers). Papers on the other classification (specifically, MCI vs AD, AD vs MCI vs HC, MCIc vs MCIreverting (MCIr), naMCI vs aMCI vs SCI, for a total of 24 studies) are shown in Table S1 (see Supplementary Materials), because not included in the meta-analysis due to the high heterogeneity of comparisons (only 4 studies compared MCI vs AD and some papers performed more than one comparison, so they could have been reported in more than one table).

Each table reports also the sample size and the follow-up duration (in terms of years) adopted to assess the conversion to Alzheimer's type dementia or to ensure a stable diagnosis (when available).

[insert Table 1, Table 2, Table 3 here]

Concerning MCI vs HC (Table 1), 20 papers were retrieved from the search. Cohort characteristics were very different among the considered studies. The median (range) of the cohort size was 60 (8-763) for MCI patients and 63 (8-5883) for HC. Six studies used subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (marked with * in the table). In these papers, measures included in the classification were extracted from neuropsychological, linguistic, demographical, anamnestic, and neuroimaging. The most frequently used ML algorithms were Support Vector Machine (SVM) and regression. The most common validation method was cross-validation.

Concerning MCIc vs MCInc (Table 2), 24 papers were included, 17 of which used subjects from the ADNI dataset. The median (range) of the cohort size was 86 (14-257) for MCIc and 100 (20-462) for MCInc. The mean of the follow up was 3 years. Measures included in the classification were neuropsychological, demographical, medical, socio-anamnestic, neuroimaging and biological data. Regression was the most frequently used ML classifier and cross-validation was the most frequently employed validation method.

Concerning AD vs HC (Table 3), 19 papers were included, 9 of which used subjects from ADNI dataset. The median value (range) of the cohort size was 59 (9-257) for AD patients and 125 (9-346) for HC. Also, for this comparison, measures included in the classification were obtained from neuropsychological, linguistic, demographical, anamnestic, neuroimaging

and biological data. The most frequently used algorithm and validation methods were SVM and cross-validation, respectively.

3.3 Risk of bias within studies

The risk of bias associated with the studies as well as the comments of the authors concerning the seven domains of the QUADAS tool has been assessed. Figures 2 and 3 show QUADAS-2 charts of the studies included in the review. Most of the studies (70%, 41/59) achieved a low *risk of bias* (Figure 2) and low *concerns regarding applicability* (Figure 3). The other papers showed a high risk of bias about the *appropriateness of reference standard* and *patient selection*.

The risk of bias tools highlighted some frequent limitations:

- only a small set of studies included sufficient details about the selection process (i.e., Cui 2012, Quintana 2012, Schmid 2013, Beltrachini 2015, Tabaton 2010, Runtti 2014);
- some studies relied on data samples of small size (i.e., less than 40 subjects overall, according to Belleville and colleagues (2017), or less than 20 subjects per diagnostic class for binary comparisons in order to perform proper training of ML algorithms); among the works included in this meta-analysis, the papers by Konig et al., (2015), Fasano et al., (2018), and Jarrold et al., (2014) fall in this category;
- most studies included in this meta-analysis did not ensure or did not provide information about the strict independence between measures used as input to the ML algorithm and measures used to assign a gold-standard diagnostic label to patients. This issue is of particular importance for studies that use neuropsychological tests. As a consequence, ML algorithms may be trained on data that are not independent of the gold-standard label to be predicted. This leads to an over-estimate of the classification performance due to circularity or double-dipping. In order to highlight this problem, some papers clearly declared that the

neuropsychological tests used in their gold standard diagnosis of patients were not used as a part of the experimental procedure (e.g. Casanova et al., 2013; Chapman et al., 2011; Gorywala et al., 2015; Konig et al., 2015; Lv et al., 2010; Quintana et al., 2012; Weakley et al., 2015; Battista 2017a);

- some studies did not have a follow up or did not specify this information in the manuscript (Lv et al., 2010; Ewers et al., 2012; García et al., 2012; Zhou et al., 2014, Goryawala et al., 2015, Konig et al., 2015, Orimaye et al., 2015; Weakley et al., 2015, Guerrero et al., 2016; Beheshti et al., 2017; Asgari et al., 2017; Fasano et al., 2018; Hernandez-Dominguez et al., 2018; Tunvirachaisakul et al., 2018);
- only one study reported, for a subsample of their patients, the post-mortem analyses and confirmation of the diagnosis (Fraser et al., 2015).

3.4 Results of the systematic review

The results of our review were obtained from the first three groups of papers (Table 1, 2 and 3), since, as previously reported, the last category (Table 1S) included a high heterogeneity of comparisons making statistical analysis not possible. Moreover, the violin plots in Figure 4 graphically show the results regarding the performance by Accuracy, Sensitivity and Specificity of the three comparisons (MCI vs HC; MCIc vs MCInc; AD vs HC). The most-frequently-used neuropsychological tests with good overall accuracy are shown in the Heatmap displayed in Figure 5. Tests are divided according to the neuropsychological domain they belong to and are ranked according to their frequency within each domain. Further, the most frequent (≥ 25% frequency) optimal predictors of the three comparisons are graphically summarized in the radar plot reported in Figure 6.

- *MCI vs HC*

Considering the set of studies that compared MCI vs HC using ML on neuropsychological data, the accuracy of classification (%) ranged from 60 to 98 (sensitivity 45-97, specificity 67-100, AUC 63-99).

Figure 5 reports the neuropsychological tests most frequently used as input for the classification and with good overall accuracy (see Appendix 1 for a full list of abbreviations of tests). Tests are divided according to the neuropsychological domain they belong to and are ranked according to their frequency within each domain. The most frequent (≥ 25% frequency) *optimal predictors* include: AVLT (43%), LM (29%), and Prose Memory Test (29%) for auditory episodic memory; MMSE (43%) for global cognitive efficiency; Category Fluency Test (36%) and BNT (29%) for language; Digit Span Test Forward and Backward Test (36%) for sustained attention and working memory; Letter Fluency Test (29%) for executive functions (Figure 6). In the Supplementary Material we reported individual tests with very good overall accuracy and/or AUC (≥ 0.7) for this diagnostic comparison (Table S3).

- *MCIc vs MCInc*

Considering the set of studies that classified MCIc vs MCInc using ML on neuropsychological tests the accuracy of classification (%) ranged from 61 to 85 (sensitivity 50-91, specificity 48-91, AUC 67-93).

Similarly, to previous results, the neuropsychological tests most frequently used as input for the classifications and with good overall accuracy are shown in Figure 5. The most frequent (≥ 25% frequency) *optimal predictors* include: AVLT (73%) and LM (33%) for auditory episodic memory; MMSE (40%) for global cognitive efficiency; TMT-B (40%) and TMT-A

(33%) for executive functions; ADAS-cog battery (33%); Digit-Span Forward and Backward test (27%) for sustained attention and working memory; Category-Fluency Test (27%) for language; FAQ (33%) for activities in daily living; GDS (27%) for depression (Figure 6). Regarding the neuropsychological tests adopted in these papers, it is interesting to note that some papers reported the total score extracted from the ADAS-cog battery (Arco et al., 2016; Casanova et al., 2013; Dukart et al., 2015; Moradi et al., 2015; Ritter et al., 2015; Runtti et al., 2014; Ye et al., 2012), while others also reported subscores of the same test, such as the Q11 sub-score (measure of word finding) (Moradi et al., 2015; Ritter et al., 2015). Behavioral and functional abilities scales were also selected in the classification (Cui et al., 2011; Dukart et al., 2015; Moradi et al., 2015; Ritter et al., 2015; Ye et al., 2012). No studies included measures of the different linguistic levels (phonological, semantic, morpho-syntactic and pragmatic) in the classification. In the Supplementary Material we reported individual tests with very good overall accuracy and/or AUC (≥ 0.7) for this comparison (Table S4).

- *AD vs HC*

Considering the studies that classified AD vs HC using ML on neuropsychological data, the accuracy of classification (%) ranged from 72 to 100 (sensitivity 73-100, specificity 77-100, AUC 79-98).

The most-frequently-used neuropsychological tests with good overall accuracy are shown in Figure 5. The most frequent (≥ 25% frequency) *optimal predictors* include: MMSE (25%) for global cognitive efficiency; AVLT (31%) for auditory episodic memory; and Category-Fluency test (38%) for language (see Appendix 1 for a full list of abbreviations of tests).

Regarding neuropsychological data, measures of verbal episodic memory were most frequently included in the final subset of neuropsychological predictors (Figure 6). Overall, measures of linguistic abilities achieved a high level of accuracy (ranging from 0.84 to 0.93),

and in particular those extracted from the picture description task (e.g., Jarrold et al., 2014; König et al., 2015). In the Supplementary Material we reported individual tests with very good overall accuracy and/or AUC (≥ 0.7) for this diagnostic comparison (Table S2).

3.5 Results of the metanalysis

Figure 7 shows the forest plots of sensitivities and specificities of the classifiers as reported in the three groups of papers (contrasts) included in the meta-analysis. Average sensitivities ranged from 73%, in the contrast MCIc vs MCInc, to 83% in the contrast MCI vs HC, up to 92% in the contrast AD vs HC, while mean specificities ranged from 69% in the contrast MCIc vs MCInc to, to 83% in the contrast MCI vs HC, up to 86% in the contrast AD vs HC.

Table 4 shows the I^2 values for the $sensitivity_m$ and $specificity_m$, and AUC_m. Consistently with the forest plots, AUC_m values were higher for AD vs HC and MCI vs HC (> 89%). However, AUC_m was good also for the MCIc vs MCInc contrast (>0.75). It should be noted that a high level of heterogeneity in $sensitivity_m$ and $specificity_m$ was found.

[insert Table 4 here]

Figure 8 completes these investigations by showing study-specific confidence regions in the ROC space and ROC curves in the contrasts of interest. The highest range of specificity was found in the contrasts AD vs HC and MCI vs HC.

Finally, we combined the data from all the neuropsychological studies to account for the effect of "comparison". The results of the random effects bivariate model are reported in Table 5. The meta-regression showed that the contrast AD vs HC has a higher sensitivity than the contrast MCI vs HC (Z = -2.12, p = 0 .034), this gap was even larger when comparing AD vs

HC with MCIc vs MCInc (-Z = 4.49, p < 0 .001). In this latter case, the contrast MCIc vs MCInc showed also a significant increment of the false positive rate (+Z = 4.91, p < 0 .001) when compared with the contrast AD vs HC.

[insert Table 5 here]

4. DISCUSSION

This is the first meta-analytic review aimed at demonstrating the reliability of ML approach trained on neuropsychological measures for performing automatic AD-related clinical screening and prognosis. Evaluating data from 59 published studies on this field of study, the majority (70%) with low risks of bias, we provided two fundamental advancements:

1) neuropsychological measures alone can lead to a successful automatic classification of prodromal AD phenotypes regardless of the employment of different ML algorithms. The contrasts MCI vs HC, MCIc vs MCInc and AD vs HC were automatically recognized with a pooled accuracy of 0.896, 0.759 and 0.914, respectively. However, the measure of heterogeneity demonstrates that the ability of ML with neuropsychological measures to predict if a patient with MCI will convert or not (the comparison between MCIc Vs MCInc) is affected by the lowest value of specificity, or in other words with a significant increment of false positive rate (Figure 4).

2) ML algorithms are able to extract relevant categories of neuropsychological tests that maximize the classification accuracy. In particular: a) MMSE, for evaluating the global cognitive status; b) AVLT, for evaluating the long-term memory performance; c) Category Fluency Test, for evaluating the language ability; and d) Digit Span Forward and Backward, for evaluating verbal short-term memory, sustained attention (I.e., TMT) and working memory capacities (Figure 6). These four neuropsychological tests showed the highest coefficient of discrimination in the ML automatic classification for all classes of interest.

Automated classification obtained with machine learning applied on neuropsychological testing: the heterogeneity question

ML algorithms can accurately detect AD and its prodromal phase but to a different extent. In particular, the results of our meta-regression approach suggest that ML algorithms have a

higher level of sensitivity in classifying AD vs HC than in classifying either MCI vs HC, or MCIc vs MCInc. A different pattern of results emerged for the false positive rate parameter, indeed while the level of false positive rate was similar between AD vs HC and MCI vs HC, ML algorithms obtained a lower level of performance in classifying MCIc vs MCInc. This pattern of results suggests that ML algorithms could support the automatic screening phase with a sensitivity higher than the 70% both with AD and MCI patients, but, at the state of the art they seem to have a relatively low prognostic power due to a relatively higher level of false positives in the contrast MCIc vs MCInc with respect to AD vs HC. However, every single paper included in the meta-analytic process has deeply been evaluated to identify the source of such heterogeneity. As expected, in most of the cases, the studies with higher variability in the estimates are those with relatively low sample sizes. For example, if one looks at the sensitivity measure for the contrast AD vs HC, the two studies with larger variability included 18 participants (Jarrold et al., 2014) and 41 participants (Konig et al., 2015), respectively. However, from Figure 7 is clear that the overall level of sensitivity heterogeneity for the contrast AD vs HC is driven by the Hernandez-Rodriguez et al (2017) study in which 257 AD patients were compared with 217 HC. As a matter of fact, if we exclude this latter study by our pooled analysis, the level of heterogeneity for sensitivity drops from 87.03 to 49.6, even though the overall AUC level does not change (from 0.91 to 0.95). On one hand, this should reassure the readers about the reliability of the results reported in this meta-analysis. On the other hand, this result actually contributes to delineating the specific route that this field of research has to follow. Indeed, our results do not suggest that smaller sample sizes are better, but that at the dawn of this field of research, relatively small sample sizes were enough to prove the concept (i.e., ML can be used also to classify patients on the basis of neuropsychological tests) and indeed, 6 out of the 11 selected papers that compared AD vs HC with a sample size smaller than 85 (with at most 41 patients) were all published between

2012 and 2016. On the contrary, studies published in the last two years adopted considerably higher sample sizes (more than 180 participants with at least 55 patients). This source of heterogeneity is intrinsic in the rapid growth of this field of research.

The best neuropsychological measures to automatically distinguish AD-related conversion

In most of the papers considered in our review, ML led to the identification of subsets of *optimal* classification features, resulting in a subset of *optimal* neuropsychological predictors that can be useful to characterize the different groups of patients (Figures 5-6). Specifically, for the classification of MCI vs HC, measures of decline in verbal episodic memory appear to be the most frequently extracted, together with measures of global cognitive status, naming, letter fluency. Concerning the classification of MCIc vs MCInc, the most frequently selected measures were verbal episodic memory, global cognitive efficiency, attentional shifting/flexibility and verbal fluency. These results are in line with the current literature, that highlights the importance of including measures not only from episodic memory, but also from more fluid functions as predictors of conversion to Alzheimer's type of dementia (Gibbons et al., 2012; Litvan et al., 2012). Two neuropsychological measures appeared to be the most commonly extracted for the classification of AD patients: verbal episodic memory and the verbal fluency tasks, i.e. measures that, from the neuropsychological point of view, tackled the most impaired functions in AD (i.e., Weintraub et al., 2012).

Of note, this meta-analytic review highlights the influence that linguistic features may have on the automated classification of Alzheimer's type dementia. Several previous works used ML algorithms to automatically extract linguistic features from the connected speech task in order to enhance classification performance (Orimaye et al., 2014; Asgari et al., 2017; Konig et al., 2015; Jarrold et al., 2014; Fraser et al., 2015; Hernandez-Rodriguez et al., 2018). Picture description tasks seem to be able to discriminate between normal and pathological cognitive

status. In addition, these measures are often analysed for neurodegeneration involving impairment in language ability as the first symptom (i.e., Primary Progressive Aphasia, PPA). Although PPA patients are out of the scope of this review, a considerable number of studies is focusing on the linguistic analysis of picture description tasks for the classification of different variants of PPA (Wilson et al., 2009; Fraser et al., 2014; Garrard et al., 2014). In brief, the results have shown that the mean length of sentences, the number of produced words and verbs, the frequency and the familiarity of words are consistent markers of PPA compared to HC. Recently, the need for a short evaluation of progressive aphasia, lead to the development of standardised language battery (Battista et al., 2017b). ML algorithms could become a useful approach to identify the combinations of language measures that most reliably and accurately classify patients based on neuropsychological/linguistic features.

Unfortunately, the evaluation of the contribution of imaging and behavioral measures with respect to biological markers was not conducted due to the paucity of studies including these data. Although the use of multimodality markers is gaining more consensus in the literature, showing that combining measures from different modalities leads to higher discriminant accuracy compared to single modality results (e.g., Westman et al., 2012), we could not perform any quantitative analysis since the number of studies providing results on multimodal approaches was poor. This is not surprising, due to the limits of biological/genetic diagnostic tests, e.g., their less readily available to clinicians that can have instead easier access to neuropsychological and neuroimaging data. In particular, among the selected papers we found only 3 studies in the category AD-HC, 5 studies in the category MCI-C and 9 studies in the category MCIc-MCInc which included details about sensitivity and specificity. In the light of the relatively sparseness of the data, we decided not to consider these studies for a formal quantitative meta-analysis, but to adopt a purely descriptive approach. From Table 1, 2 and 3 can be easily appreciated that the three studies that combined neuropsychological measures

and imaging indexes obtained a good level of sensitivity (range .84-.98) and specificity (.90-.97), while there was a higher level of variability in the results of the studies that compared MCI vs HC [Sensitivity range = .55-.97; Specificity range = .74-.94] and MCIc vs MCInc [Sensitivity range =.74-.93; Specificity range=.43-.91]. These results suggest that the combination of neuropsychological and imaging feature is a promising approach that should be better explored by future empirical studies while taking into account the methodological issues discussed in this meta-analysis to obtain more reliable and less heterogeneous performance measures that could be formally meta-analyse.

Limitations

Although in our review several neuropsychological measures were identified as the most frequently optimized measures, there are some caveats that need highlighting.

Overall, an important limitation is related to the reliability of the neuropsychological measures for the diagnosis of AD, especially in the early stages of the disease. Subjects diagnosed according to the criteria for MCI might not embody the earliest stage of AD. Our data suggest that neuropsychological tools, more sensitive than the traditional MMSE, should be taken into account to identify the earlier stages of the disease. Indeed, SCI also represents a prodromal period that could be more representative of earlier phases of the disease. Therefore, classification performance identified here should be extended further back in a phase preceding MCI in the natural history of the disease continuum.

Another relevant issue concerns the high heterogeneity among the papers selected, e.g. the neuropsychological tests used to measure the same cognitive function. For example, although the measures of the episodic memory were one of the most frequently optimized ones found in the profile of measures for AD, several tests have been reported by various authors for this measure (AVLT, HVLT, RAVLT, LM). One reason for this variety could be that there is still

no consensus as to which test leads to the best discrimination of deficits in a cognitive domain. This, along with the different ML algorithms and validation procedures associated with the accuracy of different tests, may limit the generalization of our findings.

1) Based on the quality criteria used above (QUADAS tool), some studies showed a relatively low risk of bias, while others had some features that were found to be problematic.

2) The neuropsychological measures used for clinical diagnosis of patients (i.e., the measures used to *label* patients as belonging to AD or MCI classes) can generate bias (possible over-performance) in studies using ML classifiers if the same measures are also used in the training of such classifiers (Cui et al., 2011; Kriegeskorte et al., 2009). Therefore, also the subset of best measures of progression to AD found in our review needs further investigations.

3) The scarcity of prospective longitudinal studies available for this review also represents an important limitation in the identification of neuropsychological measures for the progression of the disease. The average follow-up of the studies included in this review was around three years, and it is plausible that there was insufficient time for some patients diagnosed with MCI to progress to the clinical stage of dementia. On the other hand, it should be highlighted that there is a lack of consensus regarding how early symptoms are detected. Therefore, a longer follow up would be desirable. Eckerström et al., (2015) conducted a study enrolling MCI subjects with a ten year follow up. They found that, when considering a longer period, attention deficit as measured by TMT-B was the best measure for predicting conversion to dementia, while hippocampal volume and TMT-B attention were the best multimodal measure for conversion.

4) None of the studies used post-mortem analyses to confirm the clinical diagnosis and to assign the gold-standard diagnostic label to the patients. This may lead to methodological problems related to the possible discrepancy between clinical and definite AD diagnosis. As

highlighted above in this meta-analysis, only Fraser and colleagues (2015) used post mortem to limit this issue in a subset of the whole patients' sample.

Although the QUADAS tool allowed us recognizing and highlighting these limitations, it must be underlined that some of the items included in this tool are not appropriate to judge the quality of studies adopting ML approaches. This intrinsic issue is due to the fact that the QUADAS tool was originally conceived to assess the quality of diagnostic-accuracy tests included in systematic reviews (Whiting et al., 2011). In particular, patient-selection items seem to be unsuitable for this kind of studies. Patient selection in ML studies is usually made by selecting well-defined diagnostic groups (e.g. AD vs HC) that can be used to train a supervised classifier. This patient-selection process can be considered as a nested case-control design. The case-control design is penalized by the QUADAS tool, but it is almost necessary when analyzing data using ML techniques.

In order to overcome some of the issues described above, we identified 10 rules that could be taken into consideration by researchers in the ML domains when designing new studies using ML and neuropsychological measures for the automatic diagnosis and prognosis of AD phenotypes. These rules are reported in the Box 1 and could be useful as recommendations to design future studies.

Future works

Cognitive deficits are the last events detected in the progression of the AD disease. Unfortunately, this inevitably delays clinical diagnosis. Only one study in our review reported data on the application of neuropsychological testing and ML on subjects at a pre-clinical stage of AD (Schmid et al., 2013). This study showed that pre-clinical neuropsychological measures of AD should consider subtle qualitative decays in verbal and visual memory, visuospatial processing, error control, and subjective neuropsychological complaints. This

scarcity of studies is unexpected, since recent NIA-AA guidelines (Sperling et al., 2011; Jack et al., 2018) suggested that sensitive measures in several cognitive, functional and behavioural domains should be developed to detect early biological AD dysfunction even at a pre-clinical stage.

We would expect a focus of the more recent research on SCI, often spontaneously self-reported by the elderly, on using ML to assess the possibility that such information may represent a predictor for AD conversion even in the absence of objective deficit from the neuropsychological assessment. Since the actual predictive value for SCI remains unclear (Hohman et al., 2011; Mol et al., 2006) it would be interesting in the future to understand if using ML algorithms and longitudinal studies it might be possible to estimate this predictive value by extracting, at this very early stage, new unexpected sensitive and specific neuropsychological measures, or by increasing the specificity and sensitivity of already known neuropsychological tests by selecting a set of best predictors.

Moreover, we want to stress that future studies should also aim to evaluate whether other neuropsychological scores or sub-scores (i.e., speed, precision etc.) could be used for the classification. It must be considered that the performance might be confirmed or improved if other neuropsychological measures -not considered in the studies reported in this review- were included in the classification process. For instance, none of the studies analysed in this review evaluated the use of the Free and Cued Selective Reminding Test, developed by Dubois and colleagues (Dubois et al., 2010, 2014) for the early and differential diagnosis of AD.

Finally, very few studies compared the quality of classification and prediction based on neuropsychological tests as performed by a clinician or by an automatic classifier. Kloppel et al., (2008), showed that ML algorithms classify typical AD using MRI scans with an accuracy comparable to well-trained neuroradiologists. To the best of our knowledge, no study has been published comparing the classification of AD and MCI groups by ML with

neuropsychologist/neurologist's classification accuracy, thereby further works are necessary to disentangle this issue.

5. CONCLUSION

This meta-analytic review demonstrates that ML applied on neuropsychological measures can be useful to automatically classify AD patients, even at an early stage of the disease, and to identify a combination of optimal neuropsychological predictors. In particular, it emerged that ML and neuropsychological assessment could be used for screening purpose. This brings several advantages, such as the development of more objective and efficient neuropsychological batteries for improving the neuropsychological contribution to the early diagnosis of Alzheimer's type of dementia. Future studies in this field should empirically test the combination of methodological features necessary to improve patients' classification also at the preclinical stages.

Other information

Acknowledgements: We wish to thank Marco Piccininni for his useful comments to the manuscript.

Appendix A

Glossary of the abbreviations of Neuropsychological Tests used by the included papers

Abbreviation	Full term
ADAS	Alzheimer Disease Assessment Scale
ADL	Activities of daily living
ANT	Attention Network Test
AVLT	Auditory verbal Learning Test
BCR	Buschke's Cued Recall
BDRS	Blessed Dementia Rating Scale
BLAD	Battery of Lisbon for the Assessment of Dementia
BNT	Boston Naming Test
BVMT-R	Brief Visuospatial Memory Test-Revised
BVRT	Benton Visual Retention Test
CAMDEX	Cambridge Mental Disorders of the Elderly Examination
CDR	Clinical Dementia Rating Scale
CDR-SOB	Clinical Dementia Rating Sum of Boxes
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
CLOX	Clock drawing
CNT	Confrontation Naming Test
COWA	Controlled Oral Word Association Test
CTA	Computerized Test of Attention
CVLT	California Verbal Learning Test
DCT	Digit Cancellation Test
DLST	Digit Letter Substitution Test
DRS	Dementia Rating Scale
DS	Digit Span
DSCT	Digit Symbol-Coding Task
DSST	Digit Symbol Substitution Test
FAQ	Functional Assessment Questionnaire
FCI	Financial Capacity Instrument

Abbreviation	Full term
ADAS	Alzheimer Disease Assessment Scale
ADL	Activities of daily living
ANT	Attention Network Test
AVLT	Auditory verbal Learning Test
GDS	Geriatric Depression Scale
GPT	Grooved Pegboard Test
HVLT	Hopkins Verbal Learning Test
IADL	Instrumental Activities of daily living
LM	Logical Memory
MMSE	Mini Language State Examination
MWT	Mehrfach-Wortwahl Test
NART	North American National Adult Reading Test
Neuropsychological Bat	Neuropsychological Battery
NPI	Neuropsychiatric Inventory Questionnaire score
PFT	Phonemic Fluency Test
RAVLT	Rey Auditory verbal Learning Test
ROCF	Rey-Osterreieth Complex Figure
SDMT	Symbol Digit Memory Test
SFT	Semantic Fluency Test
SILS	Shipley Institute of Living Scale
TICS	Telephone Interview for Cognitive Status
TMT	Trial Making Test
VAT	Visual Association Test
VFT	Verbal Fluency Tests
VPAL	Verbal Paired Associates Learning
WAIS	Wechsler Adult Intelligence Scale
WMS-III	Wechsler Memory Scale
WSUI	Washington State University Instrumental

REFERENCES

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., 2011. 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 7, 270-279.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. 2017. Neuroimage, 145, 137-165.

Arco, J.E., Ramírez, J., Górriz, J.M., Puntonet, C.G., Ruz, M., 2016. Short-term Prediction of MCI to AD Conversion Based on Longitudinal MRI Analysis and Neuropsychological Tests, Innovation in Medicine and Healthcare 2015. Springer, pp. 385-394.

Asgari, M., Kaye, J., & Dodge, H. Predicting mild cognitive impairment from spontaneous spoken utterances, 2017. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 3(2), 219-228.

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

Beach TG, Monsell SE, Phillips LE, Kukull W. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. J Neuropathol Exp Neurol 2012; 71: 266–73

Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's ; disease. N Engl J Med 2012;367:795–804

Battista, P., Salvatore, C., & Castiglioni, I. (2017a). Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study. Behavioural neurology, 2017.

Battista, P., Miozzo, A., Piccininni,... Cappa S & Logroscino, G. (2017b). Primary progressive aphasia: a review of neuropsychological tests for the assessment of speech and language disorders. Aphasiology, 31(12), 1359-1378.

Beheshti, I., Maikusa, N., Matsuda, H., Demirel, H., & Anbarjafari, G. (2017). Histogram-based feature extraction from individual gray matter similarity-matrix for Alzheimer's disease classification. Journal of Alzheimer's Disease, 55(4), 1571-1582.

Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V., & Croteau, J. (2017). Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: a systematic review and meta-analysis. Neuropsychology review, 27(4), 328-353.

Beltrachini, L., Marco, M.D., A Taylor, Z., Lotjonen, J., F Frangi, A., Venneri, A., 2015. Integration of Cognitive Tests and Resting State fMRI for the Individual Identification of Mild Cognitive Impairment. Current Alzheimer research 12, 592-603.

Bishop, C.M., 2006. Pattern Recognition. Machine Learning.

Bryan RN Machine Learning Applied to Alzheimer Disease. Radiology. 2016 Dec;281(3):665-668

Casanova, R., Hsu, F.-C., Sink, K.M., Rapp, S.R., Williamson, J.D., Resnick, S.M., Espeland, M.A., Initiative, A.s.D.N., 2013. Alzheimer's disease risk assessment using large-scale machine learning methods. PLoS One 8, e77949.

Chapman, R.M., Mapstone, M., McCrary, J.W., Gardner, M.N., Porsteinsson, A., Sandoval, T.C., Guillily, M.D., DeGrush, E., Reilly, L.A., 2011. Predicting conversion from mild cognitive impairment to Alzheimer's disease using neuropsychological tests and multivariate methods. Journal of clinical and experimental neuropsychology 33, 187-199.

Clark, D., Kapur, P., Geldmacher, D., Brockington, J., Harrell, L., DeRamus, T., Blanton, P., Lokken, K., Nicholas, A., Marson, D., 2014. Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. Cortex 55, 202-218.

Craig-Schapiro, R., Kuhn, M., Xiong, C., Pickering, E.H., Liu, J., Misko, T.P., Perrin, R.J., Bales, K.R., Soares, H., Fagan, A.M., 2011. Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis. PLoS One 6, e18850.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J.S., 2011. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PLoS One 6, e21896.

Cui, Y., Wen, W., Lipnicki, D.M., Beg, M.F., Jin, J.S., Luo, S., Zhu, W., Kochan, N.A., Reppermund, S., Zhuang, L., 2012. Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach. Neuroimage 59, 1209-1217.

Doebler, P., & Holling, H. (2015). Meta-analysis of diagnostic accuracy with mada. Reterieved at: https://cran. rproject. org/web/packages/mada/vignettes/mada. Pdf)

Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., DeKosky, S.T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N.C., Galasko, D., 2010. Revising the definition of Alzheimer's disease: a new lexicon. The Lancet Neurology 9, 1118-1127.

Dubois, B., Feldman, H.H., Jacova, C., DeKosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. The Lancet Neurology 6, 734-746.

Dubois, B., Feldman, H.H., Jacova, C., Hampel, H., Molinuevo, J.L., Blennow, K., DeKosky, S.T., Gauthier, S., Selkoe, D., Bateman, R., 2014. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. The Lancet Neurology 13, 614-629.

Dukart, J., Sambataro, F., Bertolino, A., 2015. Accurate Prediction of Conversion to Alzheimer’s Disease using Imaging, Genetic, and Neuropsychological Biomarkers. Journal of Alzheimer's Disease, 1-17.

Eckerström, C., Olsson, E., Klasson, N., Berge, J., Nordlund, A., Bjerke, M., Wallin, A., 2015. Multimodal prediction of dementia with up to 10 years follow up: the Gothenburg MCI study. Journal of Alzheimer's Disease 44, 205-214.

Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L., Initiative, A.s.D.N., 2015. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiology of aging 36, S23-S31.

Ewers, M., Walsh, C., Trojanowski, J.Q., Shaw, L.M., Petersen, R.C., Jack, C.R., Feldman, H.H., Bokde, A.L., Alexander, G.E., Scheltens, P., 2012. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiology of aging 33, 1203-1214. e1202.

Fasano, F., Mitolo, M., Gardini, S., Venneri, A., Caffarra, P., & Pazzaglia, F. Combining structural magnetic resonance imaging and Visuospatial tests to classify Mild Cognitive Impairment., 2018. Current Alzheimer Research, 15(3), 237-246.

Fraser, K.C., Meltzer, J.A., Rudzicz, F., 2015. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. Journal of Alzheimer's Disease 49, 407-422.

Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM (2010): The clinical use of structural MRI in alzheimer disease. Nat Rev Neurol 6:67–77.

Gainotti, G., Quaranta, D., Vita, M.G., Marra, C., 2014. Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease. Journal of Alzheimer's Disease 38, 481-495.

García, J.M., Báez, P.G., Del Pino, M., Viadero, C.F., Araujo, C.S., 2012. A Counterpropagation Network based system for screening of Mild Cognitive Impairment, Intelligent Systems and Informatics (SISY), 2012 IEEE 10th Jubilee International Symposium on. IEEE, pp. 67-72.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., Gorno-Tempini, M.L., 2014. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex 55, 122-129.

Gibbons, L.E., Carle, A.C., Mackin, R.S., Harvey, D., Mukherjee, S., Insel, P., Curtis, S.M., Mungas, D., Crane, P.K., Initiative, A.s.D.N., 2012. A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. Brain imaging and behavior 6, 517-527.

Goryawala, M., Zhou, Q., Barker, W., Loewenstein, D.A., Duara, R., Adjouadi, M., 2015. Inclusion of neuropsychological scores in atrophy models improves diagnostic classification of alzheimer's disease and mild cognitive impairment. Computational intelligence and neuroscience 2015, 56.

Grassi, M., Perna, G., Caldirola, D., Schruers, K., Duara, R., & Loewenstein, D. A. (2018). A clinically-translatable machine learning algorithm for the prediction of alzheimer's disease conversion in individuals with mild and premild cognitive impairment. Journal of Alzheimer's Disease, 61(4), 1555-1573.

Guerrero, J., Martínez-Tomás, R., Rincón, M., Peraita, H., 2016. Diagnosis of Cognitive Impairment Compatible with Early Diagnosis of Alzheimer's Disease. Methods of information in medicine 55, 42-49.

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 260-268.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., Initiative, A.D.N., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55, 574-589.

Hohman, T.J., Beason-Held, L.L., Lamar, M., Resnick, S.M., 2011. Subjective cognitive complaints and longitudinal changes in memory and brain function. Neuropsychology 25, 125.

Jack Jr, C. R., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., ... & Phelps, C. H. (2011). Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia, 7(3), 257-262.

Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., ... & Liu, E. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14(4), 535-562.

Jack CR Jr, Holtzman DM. Biomarker modeling of Alzheimer’s disease. Neuron 2013;80:1347–58.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J., 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech, Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology, pp. 27-36.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., ... & Frackowiak, R. S. (2008). Automatic classification of MR scans in Alzheimer's disease. Brain, 131(3), 681-689.

Koikkalainen, J., Pölönen, H., Mattila, J., Van Gils, M., Soininen, H., Lötjönen, J., Initiative, A.s.D.N., 2012. Improved classification of Alzheimer's disease data via removal of nuisance variability. PLoS One 7, e31112.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 112-124.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience, 12(5), 535-540.

Lautenschlager, N.T., Flicker, L., Vasikaran, S., Leedman, P., Almeida, O.P., 2005. Subjective memory complaints with and without objective memory impairment: relationship with risk factors for dementia. The American journal of geriatric psychiatry 13, 731-734.

Litvan, I., Goldman, J.G., Tröster, A.I., Schmand, B.A., Weintraub, D., Petersen, R.C., Mollenhauer, B., Adler, C.H., Marder, K., Williams-Gray, C.H., 2012. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Movement Disorders 27, 349-356.

Lv, S., Wang, X., Cui, Y., Jin, J., Sun, Y., Tang, Y., Bai, Y., Wang, Y., Zhou, L., 2010. Application of attention network test and demographic information to detect mild cognitive impairment via combining feature selection with support vector machine. Computer Methods and programs in Biomedicine 97, 11-18.

Mateos-Pérez, J. M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., & Evans, A. C. (2018). Structural neuroimaging as clinical predictor: A review of machine learning applications. NeuroImage: Clinical.

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

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7, 263-269.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine 151, 264-269.

Mol, M.E., van Boxtel, M., Willems, D., Jolles, J., 2006. Do subjective memory complaints predict cognitive dysfunction over time? A six-year follow-up of the Maastricht Aging Study. International journal of geriatric psychiatry 21, 432-441.

Moradi, E., Hallikainen, I., Hänninen, T., Tohka, J., & Alzheimer's Disease Neuroimaging Initiative. (2017). Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease. NeuroImage: Clinical, 13, 415-427.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.s.D.N., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398-412.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., ... & Beckett, L. 2005. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia, 1(1), 55-66.

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage: Clinical 4, 687-694.

Orimaye, S.O., Tai, K.Y., Wong, J.S.-M., Wong, C.P., 2015. Learning Linguistic Biomarkers for Predicting Mild Cognitive Impairment using Compound Skip-grams. arXiv preprint arXiv:1511.02436.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience & Biobehavioral Reviews, 36(4), 1140-1152.

Pereira, T., Lemos, L., Cardoso, S., Silva, D., Rodrigues, A., Santana, I., ... & Madeira, S. C. (2017). Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. BMC medical informatics and decision making, 17(1), 110.

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

Petersen RC, Doody R, Kurz A et al. Current concepts in mild cognitive impairment. Arch Neurol 2001;58(12):1985–1992

Petersen, R.C., 2004. Mild cognitive impairment as a diagnostic entity. Journal of internal medicine 256, 183-194.

Petersen, R.C., Negash, S., 2008. Mild cognitive impairment: an overview. CNS spectrums 13, 45-53.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Archives of neurology 56, 303-308.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P., 2013. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's & Dementia 9, 63-75. e62.

Querfurth, H.W., LaFerla, F.M., 2010. Mechanisms of disease. N Engl J Med 362, 329-344.

Quintana, M., Guàrdia, J., Sánchez-Benavides, G., Aguilar, M., Molinuevo, J.L., Robles, A., Barquero, M.S., Antúnez, C., Martínez-Parra, C., Frank-García, A., 2012. Using artificial neural networks in clinical neuropsychology: High performance in mild cognitive impairment and Alzheimer's disease. Journal of clinical and experimental neuropsychology 34, 195-208.

Reverberi, C., Cherubini, P., Baldinelli, S., Luzzi, S., 2014. Semantic fluency: Cognitive basis and diagnostic performance in focal dementias and Alzheimer's. Cortex 54, e164.

Ritchie, L.J., Tuokko, H., 2010. Clinical decision trees for predicting conversion from cognitive impairment no dementia (CIND) to dementia in a longitudinal population-based study. Archives of Clinical Neuropsychology, acq089.

Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., Haynes, J.-D., Initiative, A.s.D.N., 2015. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 206-215.

Runtti, H., Mattila, J., van Gils, M., Koikkalainen, J., Soininen, H., Lötjönen, J., 2014. Quantitative evaluation of disease progression in a longitudinal mild cognitive impairment cohort. Journal of Alzheimer's Disease 39, 49-61.

Salvatore, C., Battista, P., Castiglioni, I., 2016. Frontiers for the early diagnosis of AD by means of MRI brain imaging and Support Vector Machines. Current Alzheimer research. Salvatore, C., Cerasa, A., Battista, P., Gilardi, M.C., Quattrone, A., Castiglioni, I., Initiative, A.s.D.N., 2015. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. Frontiers in neuroscience 9.

Schmid, N.S., Taylor, K.I., Foldi, N.S., Berres, M., Monsch, A.U., 2013. Neuropsychological signs of Alzheimer's disease 8 years prior to diagnosis. Journal of Alzheimer's Disease 34, 537-546.

Schrouff, J., Rosa, M.J., Rondina, J.M., Marquand, A.F., Chu, C., Ashburner, J., Phillips, C., Richiardi, J., Mourão-Miranda, J., 2013. PRoNTo: pattern recognition for neuroimaging toolbox. Neuroinformatics 11, 319-337.

Segovia, F., Bastin, C., Salmon, E., Górriz, J.M., Ramírez, J., Phillips, C., 2014. Combining PET images and neuropsychological test data for automatic diagnosis of Alzheimer's disease. PLoS One 9, e88687.

Silva, D., Guerreiro, M., Santana, I., Rodrigues, A., Cardoso, S., Maroco, J., de Mendonça, A., 2013. Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting. Journal of Alzheimer's Disease 34, 681-689.

Solfrizzi, V., Panza, F., Colacicco, A., D'introno, A., Capurso, C., Torres, F., Grigoletto, F., Maggi, S., Del Parigi, A., Reiman, E., 2004. Vascular risk factors, incidence of MCI, and rates of progression to dementia. Neurology 63, 1882-1891.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., 2011. Toward defining the pre-clinical 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 7, 280-292.

Tabaton, M., Odetti, P., Cammarata, S., Borghi, R., Monacelli, F., Caltagirone, C., Bossù, P., Buscema, M., Grossi, E., 2010. Artificial neural networks identify the predictive values of risk factors on the conversion of amnestic mild cognitive impairment. Journal of Alzheimer's Disease 19, 1035-1040.

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., 2015. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. ISCA.

Toussaint, P.-J., Perlbarg, V., Bellec, P., Desarnaud, S., Lacomblez, L., Doyon, J., Habert, M.-O., Benali, H., 2012. Resting state FDG-PET functional connectivity as an early biomarker of Alzheimer's disease using conjoint univariate and independent component analyses. Neuroimage 63, 936-946.

Tunvirachaisakul, C., Supasitthumrong, T., Tangwongchai, S., Hemrunroj, S., Chuchuen, P., Tawankanjanachot, I., ... & Maes, M. 2018. Characteristics of mild cognitive impairment using the Thai version of the Consortium to Establish a Registry for Alzheimer's Disease tests: a multivariate and machine learning study. Dementia and geriatric cognitive disorders, 45(1), 38-48.

Weakley, A., Williams, J.A., Schmitter-Edgecombe, M., Cook, D.J., 2015. Neuropsychological test selection for cognitive impairment classification: A machine learning approach. Journal of clinical and experimental neuropsychology 37, 899-916.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., ... & Petersen, R. C. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimer's & Dementia, 13(4), e1-e85.

Weintraub, S., Wicklund, A.H., Salmon, D.P., 2012. The neuropsychological profile of Alzheimer disease. Cold Spring Harbor perspectives in medicine 2, a006171.

Westman, E., Muehlboeck, J.-S., Simmons, A., 2012. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage 62, 229-238.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... & Bossuyt, P. M. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine, 155(8), 529-536.

Wilson, S.M., Ogar, J.M., Laluz, V., Growdon, M., Jang, J., Glenn, S., Miller, B.L., Weiner, M.W., Gorno-Tempini, M.L., 2009. Automated MRI-based classification of primary progressive aphasia variants. Neuroimage 47, 1558-1567.

Yang, X., Tan, M.Z., Qiu, A., 2012. CSF and brain structural imaging markers of the Alzheimer's pathological cascade. PLoS One 7, e47406.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V.A., 2012. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC neurology 12, 1.

Zhou, Q., Goryawala, M., Cabrerizo, M., Wang, J., Barker, W., Loewenstein, D.A., Duara, R., Adjouadi, M., 2014. An optimal decisional space for the classification of Alzheimer's disease and mild cognitive impairment. Biomedical Engineering, IEEE Transactions on 61, 2245-2253.

Figure Captions

Figure 1. PRISMA flow diagram depicting the different phases of the review selection process.

Figure 2. Proportion of studies included with low, high or unclear risk of bias

Figure 3. Proportion of studies included with low, high or unclear concerns regarding applicability

Figure 4. Violin plots of model performance by Accuracy, Sensitivity and Specificity stratified by the different comparisons, MCI vs HC (violet), MCIc vs MCInc (blue), and AD vs HC (green).

Figure 5. Heatmap of the neuropsychological tests most frequently used as input for the classifications and with good overall accuracy and/or AUC. Tests are divided according to the neuropsychological domain they belong to and are ranked according to their frequency within each domain. The frequency of each neuropsychological test and for each binary comparison is reported in the heatmap.

Figure 6. Radar plot of the most frequent optimal predictors (≥ 25% frequency), for the different comparisons, MCI vs HC (violet), MCIc vs MCInc (blue), and AD vs HC (green).

Figure 7. Sensitivity and Specificity Forrest Plots for the three contrasts of interest (MCI vs HC, MCIc vs MCInc, and AD vs HC).

Figure 8. Study-specific confidence regions in the ROC space and ROC curves in the contrasts of interest. The dots represent the study-specific estimates and the ellipses are obtained by plotting confidence intervals for the sensitivities and false positive rates.

Table 1 Characteristics of studies targeted MCI vs HC. The table reports the first author, year of publication and information about the use (or not) of data obtained from the ADNI public repository (marked as * if ADNI was used); the sample size; the follow up (in terms of years) adopted to assess the conversion to Alzheimer's type dementia or to ensure a stable diagnosis (when available); the modality (or modalities) of data used for the classification; the classification algorithm; the method used to validate and test the classifier; the performance of classification in terms of accuracy, specificity, sensitivity, and AUC (the best performance was reported when different classifiers were used).

Author	Sample size	Follow up (y)	Modalities of data	Classification algorithm	Validation-and-testing method	Performance [acc / sen / spe / AUC]			
						NPS	NPS+IMG	NPS+BIO	NPS+IMG+BIO
Lv et al., 2010	42 MCI; 45 HC	-	Neuropsychological; Demographical	SVM	Train-and-test	.85/.85/.86/.92			
Hinrichs et al., 2011*	119 MCI; 66 HC	2	Neuropsychological; Neuroimaging; Biological	SVM (Multi Kernel Learning)	Cross Validation	-/-/-/.74			-/-/-/.77
Cui et al., 2012*	33 MCI; 153 HC	2	Neuropsychological; Neuroimaging	SVM	Cross Validation	.64/.46/.68/.63	.79/.73/.80/.84		
Garcia et al., 2012	18 MCI; 39 HC (this dataset was extended using an over-sampling strategy)	-	Neuropsychological; Demographical	Neural Networks	Train-and-test	.91/.87/.94/-			
Quintana et al., 2012	79 MCI; 346 HC	0.5 (at least)	Neuropsychological; Demographical	Neural Networks	Train-and-test	.98/-/-/-			
Casanova et al., 2013*	153 MCIc; 182 MCInc; 188 HC	3	Neuropsychological; Neuroimaging	Logistic Regression	Nested Cross Validation	-			
						-			
Schmid et al., 2013	29 MCIc; 29 HC	~8	Neuropsychological; Anamnestic	Regression	Cross Validation	.60/-/-/-			
Zhou et al., 2014	67 aMCI; 56 naMCI; 127 HC	-	Neuropsychological; Neuroimaging	SVM	Cross Validation		.75/.61/.83/-		
							.74/.55/.82/-		
Beltrachini et al., 2015	29 MCI; 21 HC	Yes, but not specified	Neuropsychological; Neuroimaging	Linear Discriminant Analysis	Cross Validation	.93/.93/.94/.95	.96/.96/.95/.95		

				Quadratic Discriminant Analysis		.89/.88/.91/.92	.94/.92/.97/.94		
Goryawala et al., 2015*	114 early-MCI; 91 late-MCI; 125 HC	-	Neuropsychological; Demographical; Neuroimaging	Linear Discriminant Analysis	Cross Validation	.85/-/-/-	.86/.96/.74/-		
						.89/-/-/-	.91/.97/.84/-		
Konig et al., 2015	23 MCI; 15 HC	-	Neuropsychological; Linguistic	SVM	Cross Validation	.79/.79/.79/-			
Orimaye et al., 2015	Cohort I: 19 MCI; 19 HC; Cohort II: 8 MCI; 8 HC	-	Linguistic	SVM	Cross Validation	-/-/.97/.99			
Salvatore et al., 2015*	76 MCIc; 162 HC	1.5	Neuropsychological; Neuroimaging	SVM	Nested Cross Validation		.78/-/-/-		
Weakley et al., 2015	97 MCI; 161 HC	-	Neuropsychological; Demographical	Naive Bayes	Train-and-test	.92/.98/.81/-			
Asgari et al., 2017	14 MCI; 27 HC	-	Linguistic and phonetic metrics	SVM	Cross Validation	.83/.81/.76/.80			
Battista et al., 2017a*	143 MCI; 126 HC	1.5-3	Neuropsychological	SVM	Nested Cross Validation	.86/.84/.89/-			
Lin et al., 2017	763 aMCI; 253 naMCI; 127 Dementia; 5883 HC	4	Neuropsychological	SVM	Cross Validation	.71/.71/.71/.77			
						.75/.57/.75/.73			
Fasano et al., 2018	11 aMCI; 11 HC	-	Neuropsychological; Neuroimaging	SVM	Nested Cross Validation	.91/.85/1/-	1/1/1/-		
Tunvirachaisakul et al., 2018	60 MCI; 63 HC	-	Neuropsychological	SVM	Cross Validation	.77/-/-/-			

Table 2 Characteristics of studies targeted MCIc vs MCInc. The table reports the first author, year of publication and information about the use (or not) of data obtained from the ADNI public repository (marked as * if ADNI was used); the sample size; the follow up (in terms of years) adopted to assess the conversion to Alzheimer's type dementia or to ensure a stable diagnosis (when available); the modality (or modalities) of data used for the classification; the classification algorithm; the method used to validate and test the classifier; the performance of classification in terms of accuracy, specificity, sensitivity, and AUC (the best performance was reported when different classifiers were used).

Author	**Sample size**	**Follow up (y)**	**Modalities of data**	**Classification algorithm**	**Validation-and-testing method**	**Performance [acc / sen / spe / AUC]**			
						NPS	NPS+IMG	NPS+BIO	NPS+IMG+BIO
Tabaton et al., 2010	37 MCIc; 43 MCInc	2	Neuropsychological; Demographical; Biological	Neural Networks	Cross Validation			.80/.77/.83/-	
Chapman et al., 2011	Cohort I: 29 MCInc; 14 MCIc; Cohort II: 55 AD; 78 HC; 35 MCI; 5 AAMI	1.7	Neuropsychological	Discriminant Analysis	Cross Validation	.79/.79/.79/-			
Cui et al., 2011*	56 MCIc; 87 MCInc (training on 96 AD and 111 HC)	2 (at least)	Neuropsychological; Neuroimaging; Biological	SVM	Train-and-test	.65/.91/.48/.76	.62/.93/.43/.78	.65/.95/.46/.78	.67/.96/.48/.80
Hinrichs et al., 2011*	119 MCI, including MCIc, MCInc, and reverting MCI (training on 48 AD and 66 HC)	2-3	Neuropsychological; Neuroimaging; Biological	SVM (Multi Kernel Learning)	Train-and-test	-/-/-/.74			-/-/-/.77
Ye et al., 2012*	142 MCIc; 177 MCInc	4	Neuropsychological; Neuroimaging; Biological	Logistic Regression	Leave One Out	-/-/-/.77		-/-/-/.81	-/-/-/.86

Ewers et al., 2012*	58 MCIc; 72 MCInc	1.9	Neuropsychological; Demographical; Neuroimaging; Biological	Logistic Regression	Cross Validation	.65/.50/.76/-	.72/.78/.68/-	.68/.82/.57/-	.76/.88/.68/-
Koikkalainen et al., 2012*	156 MCIc; 222 MCInc	3	Neuropsychological; Demographical; Neuroimaging; Biological	Linear Regression	Cross Validation	.69/-/-/-			
Toussaint et al., 2012*	40 MCIc; 40 MCInc	2	Neuropsychological; Neuroimaging; Biological	SVM	Leave One Out	.62/.55/.70/-	.82/.85/.80/-	.68/.65/.70/-	.80/.75/.85/-
Casanova et al., 2013*	153 MCIc; 182 MCInc	3	Neuropsychological; Neuroimaging	Logistic Regression	Nested Cross Validation	.65/.58/.70/-			
Silva et al., 2013*	162 MCIc; 88 MCInc	5 (at least)	Neuropsychological	Linear Discriminant Analysis	Cross Validation	.79/.79/.80/-			
Clark et al., 2014*	44 MCIc; 36 MCInc	Up to 2 (at least 1)	Neuropsychological	Random Forest	Cross Validation	.84/.84/.83/.91			
Peters et al., 2014	18 MCIc; 22 MCInc	2	Neuropsychological; Neuroimaging	Logistic Regression	Leave One Out	.83/.72/.91/.93	.88/.83/.91/.98		
Runtti et al., 2014*	140 MCIc; 149 MCInc	2 (at least)	Neuropsychological; Neuroimaging; Biological	Linear Regression	Nested Cross Validation	.75/.74/.76/.80			.77/.82/.73/.82
Segovia et al., 2014*	26 MCIc; 20 MCInc	3	Neuropsychological; Demographical; Neuroimaging	SVM	Leave One Out	.85/.85/.85/.85	.89/.92/.85/.87		
Dukart et al., 2015*	177 MCIc; 265 MCInc (training on 144 AD and 112 HC)	2 (at least)	Neuropsychological; Neuroimaging; Biological	Naive Bayes	Train-and-test	.69/.85/.52/.72	.74/.74/.74/.77	.69/.87/.52/.72	.74/.74/.73/.77
							.74/.81/.68/.80		.74/.81/.67/.79
							.69/.78/.61/.72		.68/.76/.61/.72

Eskildsen et al., 2015*	161 MCIc; 227 MCInc (training on 194 AD and 226 HC)	3	Neuropsychological; Demographical; Neuroimaging	Linear Discriminant Analysis	Leave One Out	.61/.76/.51/.67			
Moradi et al., 2015*	164 MCIc; 100 MCInc	3 (up to 8)	Demographical (age); Neuropsychological; Neuroimaging	Low Density Separation / Random Forest	Nested Cross Validation	-/-/-/.88	.82/.87/.74/.90		
Ritter et al., 2015*	86 MCIc; 151 MCInc	3 (at least)	Demographical; Neuropsychological; Medical-Clinical; Neuroimaging; Biological	SVM	Nested Cross Validation	.72/-/-/-			.73/-/-/-
Salvatore et al., 2015*	76 MCIc; 134 MCInc	1.5	Neuropsychological; Neuroimaging	SVM	Nested Cross Validation		.60/-/-/-		
Arco et al., 2016*	73 MCIc; 61 MCInc	0.5 - 1	Neuropsychological; Neuroimaging	Linear Discriminant Analysis	Leave One Out		.74/.74/.74/.79		
Moradi et al., 2016*	164 MCIc; 100 MCInc (additional sample for training: 186 [180] AD; 226 HC; 130 [129] unknown MCI)	3 (up to 8)	Neuropsychological; Neuroimaging	Gaussian classifier	Cross Validation	.71/-/-/-	.75/-/-/-		
Pereira et al., 2017	257 MCIc; 462 MCInc	3.3 ± 2.8	Demographical; Clinical; Neuropsychological	Naive Bayes	Nested Cross Validation	-/.88/.71/.88			
					Train-and-test	-/.56/.70/.76			
Grassi et al., 2018	30 MCIc; 93 MCInc	3 (at least)	Socio-demographical; Clinical;	SVM	Cross Validation		.87/.87/.88/.91		

			Neuropsychological; Neuroimaging (MRI)						

Table 3 Characteristics of studies targeted AD vs HC. The table reports the first author, year of publication and information about the use (or not) of data obtained from the ADNI public repository (marked as * if ADNI was used); the sample size; the follow up (in terms of years) adopted to assess the conversion to Alzheimer's type dementia or to ensure a stable diagnosis (when available); the modality (or modalities) of data used for the classification; the classification algorithm; the method used to validate and test the classifier; the performance of classification in terms of accuracy, specificity, sensitivity, and AUC (the best performance was reported when different classifiers were used).

Author	**Sample size**	**Follow up (y)**	**Modalities of data**	**Classification algorithm**	**Validation-and-testing method**	**Performance [acc / sen / spe / AUC]**			
						NPS	NPS+IMG	NPS+BIO	NPS+IMG+BIO
Hinrichs et al., 2011*	48 AD; 66 HC	2	Neuropsychological; Neuroimaging; Biological	SVM (Multi Kernel Learning)	Cross Validation	.91/.89/.93/.98			.92/.87/.97/.98
Ewers et al., 2012*	81 AD; 101 HC	-	Neuropsychological; Demographical; Neuroimaging; Biological	Logistic Regression	Cross Validation	.91/.90/.91/-		.95/.92/.98/-	
Koikkalainen et al., 2012*	191 AD; 217 HC	3	Neuropsychological; Demographical; Neuroimaging; Biological	Linear Regression	Cross Validation	1/-/-/-			
Quintana et al., 2012	97 AD; 346 HC	0.5 (at least)	Neuropsychological; Demographical	Neural Networks	Train-and-test	1/-/-/-			
Touissant 2012*	40 AD; 40 HC	2	Neuropsychological; Neuroimaging; Biological	SVM	Leave One Out	1/1/1/-		1/1/1/-	1/1/1/-
Casanova et al., 2013*	171 AD; 188 HC	3	Neuropsychological; Neuroimaging	Logistic Regression	Nested Cross Validation	-			
Clark et al., 2014*	41 AD; 44 HC	Up to 2	Neuropsychological	Random Forest	Cross Validation	.94/.93/.95/.97			
Jarrold et al., 2014	9 AD; 9 HC	-	Linguistic	Neural Networks	Cross Validation	.88/.83/.90/-			

Reverberi et al., 2014	75 AD; 307 HC	1 (at least)	Neuropsychological	SVM	Leave One Out	.72/-/-/-			
Zhou et al., 2014	59 AD; 127 HC	-	Neuropsychological; Neuroimaging	SVM	Cross Validation		.92/.84/.96/-		
Fraser et al., 2015	167 AD (240 samples); 97 HC (233 samples)	4 to 9 (in some cases, post-mortem)	Linguistic	Logistic Regression	Cross Validation	.82/-/-/-			
Goryawala et al., 2015*	55 AD; 125 HC	-	Neuropsychological; Demographical; Neuroimaging	Linear Discriminant Analysis	Cross Validation	.92/-/-/-	.94/.96/.90/-		
Konig et al., 2015	26 AD; 15 HC	-	Neuropsychological; Linguistic	SVM	Cross Validation	.87/.87/.87/-			
Salvatore et al., 2015*	137 AD; 162 HC	1.5	Neuropsychological; Neuroimaging	SVM	Nested Cross Validation		.99/-/-/-		
Weakley et al., 2015	52 AD; 161 HC	-	Neuropsychological; Demographical	Naive Bayes	Training-and-testing	.99/1/.96/-			
Guerrero et al., 2016	39 AD; 42 HC	-	Neuropsychological	Bayesian Network	Leave One Out	.91/.87/.94/.96			
Battista et al., 2017a*	55 AD; 126 HC	1.5-3	Neuropsychological	SVM	Nested Cross Validation	.96/.95/.97/-			
Beheshti et al., 2017*	102 AD; 99 HC	-	Neuropsychological; Neuroimaging	SVM	Cross Validation	.85/.73/.98/.87 (FAQ)	.97/.96/.98/.97		
Hernandez-Dominguez et al., 2018	257 AD; 217 HC	-	Linguistic and phonetic metrics	SVM	Cross Validation	.79/.81/.77/.79			

Table 4. Measures of heterogeneity and AUC values for each single contrast.

Comparison	I^2 sensitivity$_m$	I^2 specificity$_m$	AUC$_m$
AD vs HC	87.03*	40.04	0.914
MCI vs HC	74.87*	78.11*	0.896
MCIc vs MCInc	9.84	71.07*	0.759

* χ^2 p-value < .05

Table 5. Meta-regression results. The table shows the estimates (together with their variability and 95% confidence intervals) and the test statistics for the Reitsma's bivariate model.

	Estimate	***Std. Error***	***95%ci.lb***	***95%ci.ub***	***z***	***Pr (>|z|)***
tsens. (Intercept)	2.374	0.23	1.91	2.83	10.11	<.001
tsens. Comparison MCI vs HC	-0.72	0.33	-1.38	0.56	-2.12	0.034*
tsens. Comparison MCIc vs MCInc	-1.33	0.29	-1.91	-0.75	-4.49	< 0.001*
tfpr. (Intercept)	-1.87	0.17	-2.22	1.53	-10.65	<.001
tfpr. Comparison MCI vs HC	0.4	0.26	-0.12	0.92	1.51	0.13
tfpr. Comparison MCIc vs MCInc	0.13	0.23	0.67	1.58	4.91	<.001*

BOX- 1

RECOMMENDATIONS TO DESIGN MACHINE LEARNING STUDIES FOR THE NEUROPSYCHOLOGICAL ASSESSMENT OF ALZHEIMER'S TYPE DEMENTIA

1. Provide the risks of bias of your study, as this can help you to improve the study quality.
2. Focus your research on clinical questions of current interest. To date, the most critical ML classification task is the discrimination of MCIc vs MCInc.
3. Use post-mortem analysis as gold-standard diagnosis of classes (ML supervised labels) whenever possible. If not, use only currently accepted diagnostic criteria to assign a clinical diagnosis to classes. In this last case, prefer clinical follow-up periods that are as long as possible in order to effectively assess the conversion to Alzheimer's type dementia or to ensure as-stable-as-possible clinical diagnoses over time.
4. Provide appropriate and complete information about the patient-selection process, e.g. specifying if the sample enrollment was consecutive or random, if the study was observational or cross sectional.
5. Use large-enough samples of patients. The sample size should be of -at least- 20 subjects per class (i.e., 40 subjects for binary comparisons). Moreover, balance the number of subjects among the classes.
6. Ensure a complete independence among those neuropsychological measures used for the ML classification and those used to assign the gold-standard diagnosis to the patients.
7. Provide appropriate and complete information about the study design, including the approach used to validate and test the ML classifier.
8. Ensure a complete independence among the sub-samples used to train, validate and test the ML classifier, respectively. For this purpose, adopt a *nested cross-validation* approach, whenever possible. Also data pre-processing has to be performed independently for these sub-samples.
9. Always include accuracy, sensitivity, specificity and AUC (as performance-evaluation metrics for the ML classifier). Report also other evaluation metrics when specific features (e.g. geometric mean or dominance for imbalanced-domain problems) have been assessed in you study.
10. Fully report all cognitive measures with predictive roles, not only the more significant ones, as these could be useful, in the future, to address between-studies consistency through meta-analytic methods.

Table S1 Characteristics of studies targeting other comparisons with respect to those included in the metanalysis. The table reports the first author, year of publication and information about the use (or not) of data obtained from the ADNI public repository (marked as * if ADNI was used); the comparison; the sample size; the follow up (in terms of years) adopted to assess the conversion to Alzheimer's type dementia or to ensure a stable diagnosis (when available); the modality (or modalities) of data used for the classification; the classification algorithm; the method used to validate and test the classifier; the performance of classification in terms of accuracy, specificity, sensitivity, and AUC (the best performance was reported when different classifiers were used).

Author	ADNI	Comparison	Sample size	Follow up (y)	Modalities of data	Classification algorithm	Validation-and-testing method	Performance [acc / sen / spe / AUC]			
								NPS	NPS+IMG	NPS+BIO	NPS+IMG+BIO
Hinrichs et al., 2011*	Y	MCIc vs reverting MCI	119 MCI, including MCIc, MCInc, and reverting MCI (training on 48 AD and 66 HC)	2-3	Neuropsychological; Neuroimaging; Biological	Multi Kernel Learning	Train-and-test	-/-/-/.94			-/-/-/.97
Lemos et al., 2012	N	AD vs MCI	94 AD; 583 MCI	5	Neuropsychological	Decision Tree	Cross Validation	.85/.60/.89/-			
Quintana et al., 2012	N	AD vs MCI vs HC	97 AD; 79 MCI; 346 HC	-	Neuropsychological; Demographical	Neural Networks	Train-and-test	.67/-/-/-			
Williams et al., 2013	N	AD vs MCI vs HC	53 AD; 97 MCI; 161 HC	-	Neuropsychological; Demographical	Naive Bayes	Cross Validation	.83/-/-/-			
Jarrold et al., 2014	N	AD vs HC vs FTD	9 AD; 9 HC; 30 FDT	-	Linguistic	Neural Networks	Cross Validation	.80/-/-/-			
		AD vs HC vs FTD subtypes	9 AD; 9 HC; 30 FDT					.61/-/-/-			
		AD vs FTD	9 AD; 30 FDT					.88/.58/.77/-			
Yin et al., 2014	N	AD vs MCI vs HC	167 AD; 189 MCI; 144 HC	-	Neuropsychological	Bayesian Network	Cross Validation	.85/-/-/-			
Goryawala et al., 2015*	Y	AD vs early-MCI	55 AD; 91 lMCI; 114 eMCI; 125 HC	-	Neuropsychological; Demographical; Neuroimaging	Linear Discriminant Analysis	Cross Validation	.93/-/-/-	.95/.95/.95/-		
		AD vs late-MCI						.90/-/-/-	.90/.91/.89/-		
		early-MCI vs late-MCI						.63/-/-/-	.74/.74/.73/-		
Hall et al., 2015	N	naMCI vs aMCI vs SCI	196 naMCI; 348 aMCI; 231 SCI	2.5	Neuropsychological; Neuroimaging; Biological	Linear Regression	Leave One Out	.72/-/-/-			.75/.76/.75/.83
Konig et al., 2015	N	AD vs MCI	23 MCI; 15 HC	-	Linguistic features	SVM	Cross Validation	.80/.80/.80/-			
Weakley et al., 2015	N	AD vs MCI vs HC	52 AD; 97 MCI; 161 HC	-	Neuropsychological; Demographical	Logistic Regression	Training-and-testing	.88/-/-/-			

		AD vs MCI				Logistic Regression		.88/-/-/-			
		CDR = 0 vs CDR = 0.5	25 CDR = {1;2}; 93 CDR = 0.5; 154 CDR = 0			Naive Bayes		.82/.65/.92/-			
		CDR = 0.5 vs CDR = {1;2}				Decision Tree		.94/-/-/-			
		CDR = 0 vs CDR = {1;2}				Logistic Regression		.99/.94/.99/-			
		CDR = 0 vs CDR = 0.5 vs CDR = {1;2}				Decision Tree		.81/-/-/-			
Battista et al., 2017*	Y	Severe vs Mild Impairment	55 Severe Impairment; 143 Mild Impairment	1.5-3	Neuropsychological	SVM	Nested Cross Validation	.69/.67/.70/-			
Gurevich et al., 2017	N	MCIc to AD vs non-AD MCI	70 MCIc to AD; 88 non-AD MCI	Post mortem	Neuropsychological	SVM	Leave-One-Out Cross Validation	.82/.77/.85/-			
Lin et al., 2017	N	(MCI + Dementia) vs HC	763 aMCI; 253 naMCI; 127 Dementia; 5883 HC	4	Neuropsychological	SVM	Cross Validation	.71/.68/.72/.76			
Amoroso et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	Random Forest and Deep Neural Network	Cross Validation + Testing on an independent cohort		.55/-/-/-		
Dimitriadis et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	Random Forest	Cross Validation + Testing on an independent cohort		.62/-/-/-		
Donnelly-Kehoe et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	Random Forest	Cross Validation + Testing on an independent cohort		.54/-/-/-		
Hernandez-Dominguez et al., 2018	N	HC vs AD+MCI	257 AD; 217 HC; 43 MCI	-	Linguistic and phonetic metrics	SVM	Cross Validation	.78/.85/.68/.76			
Choi and Jin, 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training:	Up to 10	Demographical (age and gender);	SVM	Cross Validation + Testing on an		.53/-/-/-		

			60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC		Neuropsychological (MMSE total score); Neuroimaging		independent cohort				
Nanni et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	Ensemble of classifiers	Cross Validation + Testing on an independent cohort		.55/-/-/-		
Ramirez et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	Random Forest	Cross Validation + Testing on an independent cohort		.56/-/-/-		
Salvatore et al., 2018a*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	SVM	Cross Validation + Testing on an independent cohort		.55/-/-/-		
Salvatore et al., 2018b*	Y	(AD + MCIc) vs (MCInc + HC)	50 AD; 50 MCIc; 50 MCInc; 50 HC	2	Neuropsychological; Neuroimaging	SVM	Nested Cross Validation		.85/.83/.87/-		
Sorensen et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	SVM	Cross Validation + Testing on an independent cohort		.55/-/-/-		
		AD vs MCI vs HC				SVM			.69/-/-/-		
Yao et al., 2018*	Y	AD vs MCIc vs MCInc vs HC	400, divided as follows Training: 60 AD; 60 MCc; 60 MCInc; 60 HC Testing: 40 AD; 40 MCIc; 40 MCInc; 40 HC	Up to 10	Demographical (age and gender); Neuropsychological (MMSE total score); Neuroimaging	Different classifiers (XGBoost and SVM for the final classification step)	Leave-One-Out Cross Validation + Testing on an independent cohort		.54/-/-/-		

Table S2 Individual tests with very good overall accuracy and/or AUC (≥ 0.7) for the diagnostic comparison of MCI vs. HC. Tests were grouped by cognitive, behavioural and functional domains and reported in decreasing order according to their selection frequency.

Test	Frequency
Global Cognitive Efficiency	
MMSE (Garcia 2012, Beltrachini 2015, Goryawala 2015, Battista 2017, Fasano 2018, Tunvirachaisakul 2018)	0.43
TICS (Weakley 2015)	0.07
General intelligence	
ANART (Hinrichs 2011, Battista 2017)	0.14
TIB (Test di Intelligenza Breve) (Fasano 2018)	0.07
WAIS – Vocabulary Test (Fasano 2018)	0.07
Auditory Episodic Memory	
AVLT (Hinrichs 2011, Beltrachini 2015, Goryawala 2015, Weakley 2015, Battista 2017, Fasano 2018)	0.43
Logical Memory Test (Quintana 2012, Orimaye 2015, Battista 2017, Tunvirachaisakul 2018)	0.29
Prose Memory Test (Quintana 2012, Beltrachini 2015, Battista 2017, Fasano 2018)	0.29
7/24 (Weakley 2015)	0.07
Memory Assessment Scales (Weakley 2015)	0.07
Paired Associates Test (Beltrachini 2015)	0.07
Verbal Semantic Encoding and Recognition (Fasano 2018)	0.07
Visual Memory	
Brief Visual Memory Test (Quintana 2012, Weakley 2015)	0.14
Rey-Osterrieth Complex Figure – Recall (Beltrachini 2015, Fasano 2018)	0.14
Visual Supraspan Test (Beltrachini 2015, Fasano 2018)	0.14
Language	
Category Fluency Test (Hinrichs 2011, Quintana 2012, Beltrachini 2015, Battista 2017, Fasano 2018)	0.36
BNT (Hinrichs 2011, Weakley 2015, Battista 2017, Fasano 2018)	0.29
Confrontation Naming Test (Quintana 2012, Beltrachini 2015)	0.14
Image Description task (Konig 2015, Orimaye 2015)	0.14
Category Words Fluency Test	0.07

(Fasano 2018)	
Counting Backwards task (Konig 2015)	0.07
Sentence Repeating task (Konig 2015)	0.07
Token Test (Beltrachini 2015)	0.07
Verbal Associative Fluency Test (Fasano 2018)	0.07
Executive Functions	
Letter Fluency Test (Beltrachini 2015, Konig 2015, Fasano 2018, Tunvirachaisakul 2018)	0.29
Similarities Test (Quintana 2012, Beltrachini 2015, Fasano 2018)	0.21
Stroop Test (Quintana 2012, Fasano 2018; time interference effect and error interference effect: Beltrachini 2015)	0.21
TMT-B (Hinrichs 2011, Weakley 2015, Battista 2017)	0.21
Raven Progressive Matrices (Beltrachini 2015, Fasano 2018)	0.14
TMT-A (Hinrichs 2011, Battista 2017)	0.14
D-KEFS (verbal fluency subtest: Weakley 2015)	0.07
Digit Cancellation Test (Beltrachini 2015)	0.07
Digit Symbol Substitution Test (Quintana 2012)	0.07
Dual Task (Fasano 2018)	0.07
Symbol Digit Modalities Test (oral and written subtests) (Weakley 2015)	0.07
Tower of London (Fasano 2018)	0.07
Wisconsin Card Sorting Test (Fasano 2018)	0.07
Sustained Attention and Working Memory	
Digit Span Test (Forward and backward: Hinrichs 2011, Quintana 2012, Beltrachini 2015, Battista 2017; backward: Fasano 2018)	0.36
Corsi Block Tapping Test (Beltrachini 2015, Fasano 2018)	0.14
ANT – Experimental Task (Lv 2010)	0.07
Multiple Feature Target Cancellation (Fasano 2018)	0.07
WAIS-III Letter-Number Span and Sequencing (Weakley 2015)	0.07
Visuo-Spatial Ability	
Clox 1 (Weakley 2015, Battista 2017)	0.14

Rey-Osterrieth Complex Figure – Copy (Beltrachini 2015, Fasano 2018)	0.14
Clock Test (Battista 2017)	0.07
Clox 2 (Weakley 2015)	0.07
Mental Rotation Test (Fasano 2018)	0.07
Visual Object and Space Perception Battery (Fasano 2018)	0.07
Behavioural Scales	
GDS (Weakley 2015, Battista 2017)	0.14
Batteries	
ADAS-cog (Battista 2017)	0.07
Activities in Daily Living	
FAQ (Battista 2017, Lin 2017)	0.14
Instrumental Activities of Daily Living (Beltrachini 2015, Weakley 2015)	0.14
Barthel's Index (Garcia 2012)	0.07
Physical Self Maintenance Scale (Beltrachini 2015)	0.07
Shipley Institute of Living Scale (Weakley 2015)	0.07

Table S3 Individual tests with very good overall accuracy and/or AUC (≥ 0.7) for the diagnostic comparison of MCIc vs. MCInc. Tests were grouped by cognitive, behavioural and functional domains and reported in decreasing order according to their selection frequency.

Test	Frequency
Global Cognitive Efficiency	
MMSE (Chapman 2011, Runtti 2014, Segovia 2014, Ritter 2015, Moradi 2015, Dukart 2015)	0.40
General intelligence	
ANART (Hinrichs 2011, Runtti 2014, Ritter 2015)	0.20
Auditory Episodic Memory	
AVLT (Chapman 2011, Cui 2011, Hinrichs 2011, Ewers 2012, Peters 2014, Runtti 2014, Segovia 2014, Dukart 2015, Moradi 2015, Moradi 2016, Pereira 2017)	0.73
Logical Memory Test (Chapman 2011, Cui 2011, Runtti 2014, Ritter 2015, Pereira 2017)	0.33
Paired Associates Test (Peters 2014)	0.07
Verbal Semantic Encoding and Recognition (Peters 2014)	0.07
Visual Memory	
Brief Visual Memory Test (Chapman 2011)	0.07
Rey-Osterrieth Complex Figure – Recall (Chapman 2011)	0.07
Language	
Category Fluency Test (Chapman 2011, Hinrichs 2011, Clark 2014, Runtti 2014)	0.27
BNT (Hinrichs 2011, Runtti 2014, Ritter 2015)	0.20
Token Test (Pereira 2017)	0.07
Executive Functions	
TMT-B (Chapman 2011, Hinrichs 2011, Ewers 2012, Peters 2014, Runtti 2014, Pereira 2017)	0.40
TMT-A (Chapman 2011, Hinrichs 2011, Peters 2014, Runtti 2014, Pereira 2017)	0.33
Digit Symbol Substitution Test (Runtti 2014, Ritter 2015)	0.13
Letter Fluency Test (Segovia 2014, Pereira 2017)	0.13
Raven Progressive Matrices (Pereira 2017)	0.07
Stroop Test (Chapman 2011)	0.07
Tower of London (Peters 2014)	0.07

Sustained Attention and Working Memory	
Digit Span Test (Forward and backward: Hinrichs 2011, Runtti 2014, Ritter 2015, Pereira 2017)	0.27
Visuo-Spatial Ability	
Clock Test (Runtti 2014)	0.07
Rey-Osterrieth Complex Figure – Copy (Chapman 2011)	0.07
Behavioural Scales	
GDS (Silva 2013, Runtti 2014, Dukart 2015, Ritter 2015)	0.27
Neuropsychiatric Inventory Questionnaire (Runtti 2014, Ritter 2015)	0.13
Batteries	
ADAS-cog (Ye 2012, Runtti 2014, Dukart 2015, Moradi 2015, Ritter 2015)	0.33
Staging Dementia	
CDR (Runtti 2014, Ritter 2015, Moradi 2015)	0.20
Activities in Daily Living	
FAQ (Cui 2011, Runtti 2014, Dukart 2015, Moradi 2015, Ritter 2015)	0.33

Table S4 Individual tests with very good overall accuracy and/or AUC (≥ 0.7) for the diagnostic comparison of AD vs. HC. Tests were grouped by cognitive, behavioural and functional domains and reported in decreasing order according to their selection frequency.

Test	Frequency
Global Cognitive Efficiency	
MMSE (Koikkalainen 2012, Touissant 2012, Goryawala 2015, Battista 2017)	0.25
TICS (Weakley 2015)	0.06
General intelligence	
ANART (Hinrichs 2011, Battista 2017)	0.13
Auditory Episodic Memory	
AVLT (Hinrichs 2011, Ewers 2012, Goryawala 2015, Weakley 2015, Battista 2017)	0.31
Logical Memory Test (Quintana 2012, Battista 2017)	0.13
Prose Memory Test (Quintana 2012, Battista 2017)	0.13
7/24 (Weakley 2015)	0.06
Memory Assessment Scales (Weakley 2015)	0.06
Visual Memory	
Brief Visual Memory Test (Quintana 2012, Weakley 2015)	0.13
Language	
Category Fluency Test (Hinrichs 2011, Ewers 2012, Quintana 2012, Reverberi 2014, Guerrero 2016, Battista 2017)	0.38
BNT (Hinrichs 2011, Weakley 2015, Battista 2017)	0.19
Image Description task (Fraser 2015, Konig 2015, Hernandez-Dominguez 2018)	0.19
Spontaneous Speech (Jarrold 2014, Fraser 2015)	0.13
Confrontation Naming Test (Quintana 2012)	0.06
Counting Backwards task (Konig 2015)	0.06
Sentence Repeating task (Konig 2015)	0.06
Executive Functions	
TMT-B (Hinrichs 2011, Weakley 2015, Battista 2017)	0.19
Letter Fluency Test (Clark 2014, Konig 2015)	0.13
TMT-A (Hinrichs 2011, Battista 2017)	0.13

D-KEFS (Weakley 2015)	0.06
Digit Symbol Substitution Test (Quintana 2012)	0.06
Similarities Test (Quintana 2012)	0.06
Stroop Test (Quintana 2012)	0.06
Symbol Digit Modalities Test (oral and written subtests) (Weakley 2015)	0.06
Sustained Attention and Working Memory	
Digit Span Test (Forward and backward: Hinrichs 2011, Quintana 2012, Battista 2017)	0.19
WAIS-III Letter-Number Span and Sequencing (Weakley 2015)	0.06
Visuo-Spatial Ability	
Clox 1 (Weakley 2015, Battista 2017)	0.13
Clock Test (Battista 2017)	0.06
Clox 2 (Weakley 2015)	0.06
Behavioral Scales	
GDS (Weakley 2015, Battista 2017)	0.13
Batteries	
ADAS-cog (Koikkalainen 2012, Touissant 2012, Battista 2017)	0.19
Staging Dementia	
CDR (Koikkalainen 2012)	0.06
Activities in Daily Living	
FAQ (Koikkalainen 2012, Battista 2017, Behesti 2017)	0.19
Instrumental Activities of Daily Living (Weakley 2015)	0.06
Shipley Institute of Living Scale (Weakley 2015)	0.06

MATERIAL AND METHODS

Information Sources and Search

Two of the authors (PB, CS) independently conducted an extensive literature search in MEDLINE using Pubmed and EMBASE using Ovid, PsychINFO and Web of Science. The search was concluded on July 15, 2018. The search strategy based on the PICOS approach was applied following five concepts: 1) Patient, defined as elderly with MCI or AD; 2) Intervention, defined as the index tests, specifically the cognitive measures used as predictors; 3) Comparison, defined as the clinical diagnosis of AD; 4) Outcome, defined as the predicted outcome, which was, for example, “conversion (or not) to Alzheimer’s type dementia”, and 5) Type of the study, which should be “longitudinal studies” or “nested case-control studies”. The following keywords (with both extended names and abbreviations) were used for the literature search: ((Alzheimer OR "mild cognitive impairment" OR neurodegenerative) AND ("neuropsychological assessment" OR "neuropsychological measure" OR "neuropsychological test" OR "cognitive assessment" OR "cognitive measure" OR "cognitive test") AND ("machine learning" OR "artificial intelligence" OR classification)). In order to increase the likelihood that all the potentially relevant studies were identified, further papers were included by the two authors from a manual search, starting from the lists of references of previously retrieved articles.

Study selection

The study selection was carried out by two reviewers (PB, CS), independently. The studies retrieved by the search strategy were first screened based on the titles and then selected by one of the two reviewers (PB) based on abstracts. One additional reviewer (AC) independently revised the list of potential articles based on abstracts. The full text of the articles considered to be potentially eligible was then evaluated in detail by the same reviewer for quality assessment and any unresolved issues were discussed with IC. All articles reporting data that could be appropriately pooled were included in the quantitative analysis. Specifically, we restricted our analysis to those papers that reported at least one measure of the automatic-classification performance among accuracy, sensitivity, specificity, and Area Under the ROC Curve (AUC).

REFERENCES

Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., ... & Alzheimer's Disease Neuroimaging Initiative. (2018). Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge. Journal of neuroscience methods, 302, 3-9.

Choi, H., Jin, K. H., & Alzheimer's Disease Neuroimaging Initiative. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behavioural brain research, 344, 103-109.

Dimitriadis, S. I., Liparas, D., & Alzheimer's Disease Neuroimaging Initiative. (2018). How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. Neural regeneration research, 13(6), 962.

Donnelly-Kehoe, P. A., Pascariello, G. O., Gómez, J. C., & Alzheimers Disease Neuroimaging Initiative. (2018). Looking for Alzheimer's Disease morphometric signatures using machine learning techniques. Journal of neuroscience methods, 302, 24-34.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J., 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech, Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology, pp. 27-36.

Goryawala, M., Zhou, Q., Barker, W., Loewenstein, D.A., Duara, R., Adjouadi, M., 2015. Inclusion of neuropsychological scores in atrophy models improves diagnostic classification of alzheimer's disease and mild cognitive impairment. Computational intelligence and neuroscience 2015, 56.

Gurevich, P., Stuke, H., Kastrup, A., Stuke, H., & Hildebrandt, H. (2017). Neuropsychological testing and machine learning distinguish Alzheimer's disease from other causes for cognitive impairment. Frontiers in aging neuroscience, 9, 114.

Hall, A., Mattila, J., Koikkalainen, J., Lotjonen, J., Wolz, R., Scheltens, P., ... & Minthon, L. (2015). Predicting progression from cognitive impairment to Alzheimer's disease with the Disease State Index. Current Alzheimer Research, 12(1), 69-79.

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 260-268.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 112-124.

Lin, M., Gong, P., Yang, T., Ye, J., Albin, R. L., & Dodge, H. H. (2018). Big data analytical approaches to the NACC Dataset: aiding preclinical trial enrichment. Alzheimer disease and associated disorders, 32(1), 18.

Lemos, L., Silva, D., Guerreiro, M., Santana, I., de Mendonça, A., Tomás, P., & Madeira, S. C. (2012). Discriminating Alzheimer's disease from mild cognitive impairment using neuropsychological data. *Age (M±SD)*, *70*(8.4), 73.

Nanni, L., Lumini, A., & Zaffonato, N. (2018). Ensemble based on static classifier selection for automated diagnosis of mild cognitive impairment. Journal of neuroscience methods, 302, 42-46.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., Initiative, A.D.N., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55, 574-589.

Quintana, M., Guàrdia, J., Sánchez-Benavides, G., Aguilar, M., Molinuevo, J.L., Robles, A., Barquero, M.S., Antúnez, C., Martínez-Parra, C., Frank-García, A., 2012. Using artificial neural networks in clinical neuropsychology: High performance in mild cognitive impairment and Alzheimer's disease. Journal of clinical and experimental neuropsychology 34, 195-208.

Ramírez, J., Górriz, J. M., Ortiz, A., Martínez-Murcia, F. J., Segovia, F., Salas-Gonzalez, D., ... & Alzheimer's Disease Neuroimaging Initiative. (2018). Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. Journal of neuroscience methods, 302, 47-57.

Salvatore, C., & Castiglioni, I. (2018). A wrapped multi-label classifier for the automatic diagnosis and prognosis of Alzheimer's disease. Journal of neuroscience methods, 302, 58-65.

Salvatore, C., Cerasa, A., & Castiglioni, I. (2018). MRI characterizes the progressive course of AD and predicts conversion to Alzheimer's dementia twenty-four months before probable diagnosis. Frontiers in aging neuroscience, 10, 135.

Sørensen, L., Nielsen, M., & Alzheimer's Disease Neuroimaging Initiative. (2018). Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. Journal of neuroscience methods, 302, 66-74.

Weakley, A., Williams, J.A., Schmitter-Edgecombe, M., Cook, D.J., 2015. Neuropsychological test selection for cognitive impairment classification: A machine learning approach. Journal of clinical and experimental neuropsychology 37, 899-916.

Williams, J. A., Weakley, A., Cook, D. J., & Schmitter-Edgecombe, M. (2013). Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In Workshops at the twenty-seventh AAAI conference on artificial intelligence.

Yao, D., Calhoun, V. D., Fu, Z., Du, Y., & Sui, J. (2018). An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment. Journal of neuroscience methods, 302, 75-81.

Yin, Z., Zhao, Y., Lu, X., & Duan, H. (2015). A hybrid intelligent diagnosis approach for quick screening of Alzheimer's disease based on multiple neuropsychological rating scales. Computational and mathematical methods in medicine, 2015.