Clinical validity of delayed recall tests as a gateway-biomarker for Alzheimer's disease in the context of a structured 5-phase development framework

Authors: Chiara Cerami ^{a,b,c}, Bruno Dubois ^d, Marina Boccardi ^{e,f}, Andreas U. Monsch ^g, Jean Francois Demonet ^h, Stefano F. Cappa ^{b,i}, and the Geneva Task Force for the Roadmap of Alzheimer's Biomarkers.

- a. Vita-Salute San Raffaele University, Milan, Italy
- b. Division of Neuroscience, San Raffaele Scientific Institute, Milan, Italy
- c. Clinical Neuroscience Department, San Raffaele Turro Hospital, Milan, Italy
- d. Dementia Research Center and Department of Neurology, Salpêtrière University Hospital, Paris University, Paris, France
- e. Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS Centro S. Giovanni di Dio Fatebenefratelli, Brescia, Italy
- f. LANVIE - Laboratory of Neuroimaging of Aging, University of Geneva, Geneva, Switzerland
- g. Memory Clinic, University Center for Medicine of Aging, Felix Platter Hospital, Basel, Switzerland
- h. Leenaards Memory Centre – CHUV, Lausanne University, Lausanne, Switzerland
- i. NeTS Center, Istituto Universitario di Studi Superiori, Pavia, Italy

Character and Word Counts: title (149 characters); abstract (174 words); text (8800 words)

References, Tables and Figures: 123 references; 1 table

Key-words: Alzheimer's Disease; Diagnosis; Biomarkers; Episodic Memory; Delayed recall; Free word list recall tests; Cued selective reminding task.

Corresponding Author: Chiara Cerami, MD

Università Vita-Salute San Raffaele and San Raffaele Scientific Institute

Division of Neuroscience

Via Olgettina, 60 – 20132 Milan (Italy)

Tel. +39 0226435760; Fax. +39 0226435738

E-mail: cerami.chiara@hsr.it

Abstract

Although Alzheimer's disease (AD) criteria promote the use of biomarkers, their maturity in clinical routine still needs to be assessed. In the light of the oncology framework, we conducted a literature review on measures used to assess delayed recall impairment due to medial temporal lobe dysfunction (i.e., free and cued word list recall tests). Ample evidence is available for phase 1 (rationale for use), 2 (discriminative ability) and 3 (early detection ability) for many of the tests in routine use. Evidence about Phase 4 (performance in real world) and phase 5 (quantify impact and costs) are yet to come. Administration procedures have been standardized and cut-off scores are well validated in large AD and MCI series. Some aspects (e.g., different task formats), however, hamper the comparability of results among different populations and the reproducibility between laboratories. No definite guideline for their use can thus be proposed at the moment. Accordingly, the maturity of such markers is not yet sufficient, and requires future investigation to promote the proper use of memory measures in clinical settings.

Keywords: episodic memory; delayed recall; cued recall tasks; free word list tests; biomarker development; early diagnosis; biomarker-based diagnosis; Alzheimer's disease; Mild Cognitive Impairment; 5-phases.

Graphical abstract

Development of **Free and Cued Wordlist Recall** adapted from the framework of Pepe et al. 2001

Phase 1: Rational for the use of Free and Cued Wordlist Recall		Phase 2: Discrimination ability of Free and Cued Wordlist Recall		Phase 3: Detection ability in early phase		Phase 4: Free and Cued Wordlist Recall accuracy in representative MCI patients		Phase 5: Quantify impact of Free and Cued Wordlist Recall-based diagnosis on relevant outcomes	
Primary aim	Primary aim	Secondary aims	Primary aims	Secondary aims	Primary aim	Secondary ais	Primary aim	Secondary aims	
Potential leads	Identify discrimination accuracy AD/HC	Assay definition	Assess capacity of earliest (MCI) detection	Impact of covariates	Assess true/false referral rate in Free and Cued Wordlist Recall diagnosed patients	Detect predictive features	Estimate impact on morbidity & disability	Cost/ benefit quantification	
		Ante mortem/ autopsy		Compare markers		Practical feasibility		Compliance in different settings	
						Estimate impact & costs		Compare different protocols	
<div>Achievement<div><div>Full</div><div>Partial</div><div>Preliminary</div><div>Not achieved</div><div>Not applicable</div></div></div>		Covariates in HC	Criteria for positivity	Combine markers					
		Covariates in AD		Determine testing interval		Monitor false negatives			

1. Introduction

The correct identification of Alzheimer's disease (AD) represents a challenge for clinicians especially at the predementia stages. Recent developments in this area of research are specifically devoted to support the early identification of AD pathology *in vivo* and to the application of reliable biomarkers of disease in clinical settings. The need of more accurate early and differential diagnosis, indeed, prompted the development of new research criteria supporting the use of biomarkers in order to recognize AD in prodromal or even preclinical stages (Dubois et al., 2007, 2010, 2014; McKhann et al., 2011; Sperling et al., 2011; Jack et al., 2011; Albert et al., 2011).

Since the introduction of new clinical criteria, AD research has been mainly focused on the application of distinctive topographic (e.g., FDG-PET; MRI) and pathophysiological (e.g. amyloid-PET or CSF) biomarkers in clinical research. Several methodological problems have emerged about their implementation in clinical routine. Neither a definite diagnostic algorithm nor clear quantitative measures for the use of biomarkers in patients suspected for AD have been clearly outlined. Thus, different biomarkers and measurement tools are used by researchers according to their availability in community-based and clinical-based studies, obviously resulting in heterogeneous findings.

To overcome similar problems in the field of oncology, Pepe and colleagues (Pepe et al., 2001) suggested systematizing the investigation of cancer biomarkers on the basis of the methodology used for pharmacological investigation. Since a formal structure to guide the process of AD biomarker development was lacking so far, an effort has recently been launched to adopt the above-mentioned oncology model to effectively systematize the available scientific evidence for the use of biomarkers in AD diagnosis, with the aim to promote rigor in their application to clinical settings. The present study focuses on the analysis of the maturity of the assessment of episodic memory by means of delayed recall tasks in the framework of this model.

Whether cognitive testing can properly be considered as **a marker of disease** is an open question. The concept is fully compatible with a broad definition, such as "...a characteristic which can be objectively measured and evaluated as an indicator of a physiological as well as a pathological process or pharmacological response to a therapeutic intervention" (Jain, 2012). In the field of dementia, however, neuropsychological testing is generally considered separately from imaging and CSF biomarkers (see, for example, Ewers et al., 2012). Within the MCI/prodromal AD context, neuropsychological testing is usually considered as a sort of "gatekeeper" for the application of biomarkers, as the presence of objective impairment is required by the diagnostic criteria to separate these conditions from subjective complaints. In any case, given the central role of cognitive assessment, it is surprising that studies assessing the sensitivity and specificity of neuropsychological tests for diagnosis of AD or the predictive value of the progression from mild memory impairment (MCI) to AD are relatively scant and heterogeneous in methodology. In particular, the presence of an early and significant objective deficit of memory and learning has been considered as the main criterion supporting the diagnosis of typical AD condition for decades (Mc Khann et al., 1984; DSM-IV, APA, 2000), and an impaired memory performance in comparison to a healthy control group is considered as the best cognitive predictor of the development of future AD (Elias et al., 2000; Small et al., 2000; Sarazin et al., 2007). According to Braak and Braak staging of AD (Braak and Braak, 1991), the earliest neuropathological changes involve the entorhinal cortex and the hippocampal structures, disconnecting the Papez circuit and selectively affecting the ability to consolidate new information. This results in an impaired performance on delayed recall memory tasks (Squire et al., 2004).

Moreover, in the context of recent diagnostic criteria for the very early/prodromal stages of AD, the performance in specific memory tests has gained a special status. The presence of impaired memory performance on objective testing is required for the definition of MCI (hence 'amnesic MCI') or prodromal AD status. If a subject presents with subjective memory complaints, in the absence of objective memory dysfunction, he/she is defined as having a "subjective memory

impairment”, an inconsistently defined construct (Abdulrab and Heun, 2008), which in clinical practice does not usually lead to further investigation, but only to reassurance and long-term follow-up (Berrios et al., 2000; Jessen et al., 2014).

Since cognitive assessment remains a critical component of diagnosis in clinical and research settings, it is vital to determine the capacity of specific memory measures in detecting early disease changes and predicting disease progression, in order to recommend tests having the greatest predictive accuracy. The tasks assessing memory ability may differ with respect to the modality of stimulus presentation, the testing procedure, the structure of the to-be-remembered information or the presence of facilitators to improve encoding and recall (cued paradigms).

Many memory measures are used in clinical and research settings. The most common are measures of delayed free recall of word lists. A variety of standardized verbal learning tasks, such as the Rey Auditory Verbal Learning Test (RAVLT; Rey, 1941), the California Verbal Learning Test (CVLT, Delis et al., 1987; CVLT-II, Delis et al., 2000), and the Hopkins Verbal Learning Test (HVLT-R, Brandt and Benedict, 2001), are commonly used for clinical diagnosis and disease monitoring of AD dementia and mild cognitive impaired (MCI) patients. In addition, some word list tasks, often quicker to administer, i.e., with fewer words to learn (10-word list) and less learning trials (2 or 3), are part of neuropsychological batteries, such as the Wechsler Memory battery (<http://www.pearsonclinical.com/>), ADAS-cog (Mohs et al., 1983), CERAD (Morris et al., 1988), or MoCA (Nasreddine et al., 2005). Direct measurement of the number of items recalled at the learning trials (i.e., the immediate recall) or after a time delay (i.e., the delayed recall), as well as the difference between immediate and delayed recall (i.e., the savings) are the main measures obtained from the word list free recall tasks used to evaluate patient performances.

Other neuropsychological tests for the assessment of verbal long-term memory are logical memory (short story recall) and associative learning tasks (e.g., in the Wechsler Memory battery). Both immediate and delayed scores are obtained from story recall tests. In this case, the processing of a coherent stream of information typically benefits from intrinsic semantic organization of the

material, while the ability to self-generate organizational strategies is required for free recall of word list tasks— with the exception of the CVLT (Randolph et al., 1994). Other tests assess non-verbal memory, such as, for example, the delayed recall of Rey figure or the Cambridge Neuropsychological Test Automated Battery (CANTAB) (<http://www.cambridgecognition.com/>), testing the ability to form and remember associations between the attributes of an experience. These tasks are sensitive to the functional integrity of the medial temporal lobe (Eichenbaum and Cohen, 2001; Melzer and Constable, 2005), as performance largely depends on the ability to bind and encode arbitrary information.

Though the amnesic syndrome is typical of the onset of AD, impairments on delayed recall tasks may also be present in non AD-neurodegenerative diseases (e.g., behavioral variant of frontotemporal dementia; Hornberger et al., 2010) as well as in other conditions (e.g., vascular mild cognitive impairment-VCI, depression), characterized by cognitive deficits that can affect the learning phase or the encoding and recall processes (Dickerson and Eichenbaum, 2010). In AD the amnesic profile is typically characterized by poor learning and rapid forgetting over relatively short periods, reflecting damage to the hippocampal structures (Squire et al., 2004). To accurately detect memory impairment of the hippocampal type, the design of the test used to assess memory ability, and in particular of the learning phase of the task, is crucial. Free recall is dependent upon intact attentional processing, registration and retrieval mechanisms. An effective retrieval of the to-be-remembered information can be better achieved with an “encoding specificity”, based on the correspondence of the semantic cue during encoding and retrieval. The use of this learning technique produces efficient results in healthy subjects (Ivnik et al., 1997). In agreement to this evidence, the specific neuropsychological features of the memory impairment of AD has been stressed in the IWG research diagnostic criteria (Dubois et al., 2010, 2014) and defined as follows: “objective evidence of significantly impaired episodic memory on testing, generally consisting of a recall deficit that does not improve significantly with cueing or recognition testing after effective encoding of information has been controlled”. Test paradigms providing controlled learning

conditions (encoding) and distinguishing between the performance on free recall and the effects of cueing non-recalled items (i.e., the Grober-Buschke paradigm, Buschke, 1984) can distinguish between the encoding and storage impairments typical of AD condition and low memory performances resulting from reduced attentional resources or ineffective recall strategies (Grober and Buschke, 1987). The Free and Cued Selective Reminding Test (FCSRT; Buschke, 1984; Grober and Buschke, 1987) has thus gradually entered standard neuropsychological batteries, partially replacing traditional free recall word list tasks (e.g., RAVLT). Three FCSRT measures are mainly used to detect AD: free recall, total recall (the sum of free and cued recall) and cue efficiency (the ratio of cued recall successes to the number of cued recall attempts). The high negative predictive values of free and total recall FCSRT (Auriacombe et al., 2010; Dierckx et al., 2009) support the utility of these measures for ruling out dementia in both primary and secondary care settings. Low positive predictive values (Auriacombe et al., 2010; Dierckx et al., 2009) may limit their usefulness. This effect is, however, more restricted among patients attending memory clinics (Auriacombe et al., 2010), where disease prevalence is higher.

For the considerations summarized above, among the large number of measures used in clinical practice to assess episodic memory ability, the present discussion is limited to delayed free and cued recall tests for several reasons. In particular, they are the most widely used tests in clinical and research settings (Maruta et al., 2011). Moreover, there is extensive evidence that they may represent the most sensitive measures of memory decline and that they may be relatively specific for medial temporal lobe dysfunction, i.e., to the earliest pathological changes in AD (Braak and Braak, 1991).

2. Methods

2.1 Target

This study was performed with reference to the model imported from the oncology field (Pepe et al., 2001) and adapted to the field of dementia, specifically to the aim of performing the differential

diagnosis of AD at the prodromal stage (Frisoni et al., 2015, this issue). The terms of this framework are summarized in this section. The target populations are sporadic AD dementia patients and MCI subjects, as defined below. Although studies with pathological confirmation should be considered as the gold standard to evaluate accuracy of memory measures in AD, the criteria chosen to select papers appropriate for this review included also studies in which AD was diagnosed according to the clinical criteria (at baseline or at follow-up) available at the time of the study. We considered progression to dementia at a minimum of 2-year follow-up as the gold standard for the clinical diagnosis of AD in MCI patients.

2.2 Glossary

Alzheimer's disease (AD). Alzheimer pathology consisting of brain amyloidosis and neurodegeneration, usually with medio-temporal and temporo-parietal distribution. The term is thus independent of the clinical manifestation of the disease.

Alzheimer's dementia (AD dementia). The clinical syndrome featuring both cognitive impairments and functional disability as defined by McKhann (1984). Notably, not all cases of AD dementia have AD pathology due to imperfect accuracy of purely clinical criteria.

Mild cognitive impairment (MCI). The population with cognitive impairment and no functional disability, including, besides cases with prodromal AD, cases with no neurodegenerative disorders (about 35-40%) and non-AD neurodegeneration (about 10-15%) (Bennett et al., 2002; Rowe et al., 2010; Jack et al., 2008). The MCI cases with biomarker positivity are defined as prodromal AD in the clinical criteria by Dubois et al. (2010). The focus of our review is the diagnosis of AD at the MCI stage.

Non-AD neurodegenerative diseases. The neurodegenerative disorders entering differential diagnosis considered in this context are Suspected Non-Amyloid Pathology (SNAP), hippocampal sclerosis, frontotemporal lobar degeneration (FTLD), tauopathies (e.g., progressive supranuclear

palsy and corticobasal degeneration), dementia with Lewy body (DLB) and other alpha-synucleinopathies such as multiple system atrophy.

2.3 Conceptual framework

Here, we summarize the results of the translation of the conceptual framework from oncology to the AD field described elsewhere (Frisoni et al., 2015, this issue), as steps to be covered in the systematic development of **disease markers** for a proper use in the clinical routine for the diagnosis of AD. The present review focuses on the maturity of selected neuropsychological tasks for the assessment of delayed recall relative to each step. **In reference to the model imported from the oncology field (Pepe et al., 2001), we generally referred to the neuropsychological markers and the other AD disease markers as “biomarkers”.** All aims and sub-aims are specifically addressed, and qualified as “*Fully achieved*”, “*Partly achieved*”, “*Preliminary evidences*”, “*Not achieved*” or “*Not applicable*” based on the available evidence. The evaluation terms and assessments are reported in detail in Table 1.

Phase 1. Phase 1 studies aim to identify the rationale of the biomarker, based on pathology findings, and consist of preclinical exploratory studies.

Phase 2. Phase 2 studies aim to define the ability of the biomarker to discriminate patients from healthy subjects. It focuses on defining the clinical essay allowing reliable discrimination, and in the assessment of the possible differential effects of covariates that may influence the thresholds for positivity.

Phase 3. Phase 3 studies aim to define the ability of the biomarker to detect the disease in its early phase, namely, in the prodromal stage as set for the present exercise. They consist of prospective longitudinal repository studies aimed to fine-tune the exact thresholds for positivity, and to compare the usefulness of the biomarker compared to or differentially combined with the other available biomarkers.

Phase 4. Phase 4 studies aim to estimate the accuracy and usefulness of the biomarker-based diagnosis in real world patients. It consists of prospective diagnostic studies and on subjects also undergoing treatment following the biomarker-based diagnosis. It assesses the benefit of the biomarker-based detection in terms of early diagnosis, feasibility, compliance, and provides preliminary evidence about mortality, costs, and undetected cases.

Phase 5. Phase 5 studies aim to exactly quantify the impact of the biomarker-based diagnosis on clinically meaningful outcomes and costs. It consists of case-control studies, assessing the reduction in mortality, morbidity, disability and costs as allowed by the biomarker-based diagnosis. It also aims to compare the costs of different biomarkers testing protocols and settings.

2.4 Evidence evaluation

The primary and secondary aims described in Section 2.3 were assessed as *Fully achieved*, *Partly achieved*, *Preliminary evidence*, *Not achieved* or *Not applicable*.

Fully achieved: A primary or secondary aim was assessed as *fully achieved* when its corresponding scientific evidence is available and replicated in adequately powered samples in studies without major methodological faults.

Partly achieved: A primary or secondary aim was assessed as *partly achieved* when scientific evidence is available but not yet sufficiently replicated, or samples are not adequately powered, or other significant methodological limitations can be found in the available literature.

Preliminary evidence: A primary or secondary aim was assessed as *preliminary evidence* when only preliminary evidence is available.

Not achieved: A primary or secondary aim was assessed as *not achieved* when no evidence was found and no studies are known to be ongoing at present to cover this aim.

Not applicable: A primary or secondary aim was assessed as *not applicable* when no evidence can be yield for the considered biomarker.

2.5 Papers search and selection

Papers were selected searching on the PubMed and Medline databases. Trying to apply the framework of Pepe et al. (2001) to the use of tests designed for the assessment of memory as a marker of the clinical presentation of typical AD, we identified studies targeting well-established candidates for the detection of long-term memory disorders in AD dementia patients or MCI subjects at risk for progression to AD dementia. Only papers including patients classified as MCI or AD according to validated clinical diagnostic and research criteria (McKhann et al., 1984, 2011; Petersen, 2004; Albert et al., 2011; Sperling et al., 2011; Dubois et al., 2010, 2014) were included.

The search was limited to the available evidence regarding free or cued word list recall tests. The reasons for the selection of this specific class of tasks are described in the introduction. We do not aim to rule out other tests or conceptualizations for testing memory related to early medial temporal cortex / hippocampal structures damage in AD dementia with new experimental procedures (see below). However, evidence is still lacking to support the applicability of new promising cognitive tools to explore AD-related memory disorders. Short story recall (logical memory) and paired associates (associative memory) learning tasks have been excluded from this review due to the different neuropsychological constructs compared to the verbal word list tasks. Some of these neuropsychological measures (e.g., paired associate learning and name-face learning) actually showed high sensitivity in sporadic and familial AD (see Lowndes and Savage, 2007 for a review on the usefulness of associative memory tasks). A promising paradigm, namely the visual short-term memory binding task (Parra et al., 2009), which showed very high sensitivity and specificity even in the preclinical phase of genetic AD cases (Parra et al., 2010), still needs additional evidence for the proper use in clinical settings. Other potentially interesting candidates for the early AD recognition are the DMS48 visual recognition test (Barbeau et al., 2004) and the topographical memory test described by Hartley et al. (2007), both specifically designed to detect the early pathological involvement of the entorhinal-perirhinal cortex. Low performances on the DMS48 task

may detect MCI subjects with a pattern of grey matter loss usually described in early AD (Barbeau et al., 2008).

Different strings were used to circumscribe the search specifically to the different aims/sub-aims (see Supplementary Table at [/centroalzheimer.it/public/MB/BM-Roadmap/BiomarkerRoadmap_Suppl.docx](http://centroalzheimer.it/public/MB/BM-Roadmap/BiomarkerRoadmap_Suppl.docx)). When aims were notoriously achieved (e.g., evidence that low performance on delayed recall tasks is a hallmark for typical AD), a set of reference papers or, if available, a review was selected by the authors. In cases with lacking or scanty output, more inclusive strings were admitted to extend the research. Relevant evidence from previous personal knowledge and tracking from reference papers was included. Only papers published in English and up to May 2015 were included. The final selection of articles was based on relevance, as judged by the authors.

3. Current maturity of free and cued word list recall tests

3.1 Phase 1 – Preclinical exploratory studies

The aim of Phase 1 is to identify leads for potentially useful biomarkers and to prioritize them. Evidence coming from large preclinical exploratory studies highly supports the rationale for the use of impaired performance at word list recall tests as marker of typical AD. Such tests assessing long term recall of recently learned items are often tagged as “episodic memory tests”, even though they capture only one important feature of the learning episode (i.e., what the subject remembers) (Tulving, 2002). In typical AD, early neuropathological changes occur in the perirhinal and entorhinal cortices, and in the hippocampal structures disconnecting the Papez circuit (Braak and Braak, 1991), and thus selectively affecting the ability to consolidate new information and resulting in low performances on delayed recall tasks (Squire et al., 2004). Medial temporal lobe structures have a fundamental role in the rapid formation of lasting memory information (Squire et al., 2004). Different studies have assessed the link between the neuropathological hallmarks of AD (i.e., amyloid plaques and neurofibrillary tangles) and the severity of ante-mortem cognitive impairment

(see Nelson et al., 2012 for a review). Despite a large variability among clinico-pathological correlation studies, mainly due to different study designs, research cohort characteristics, clinical-neuropsychological assessments or operational procedures for the anatomical-pathological analysis, studies with large autopsy series agreed in reporting that *post mortem* density of neocortical tau-positive neurofibrillary tangles is the pathological feature best correlating with *ante mortem* cognitive status (Nelson et al., 2012). In particular, the neuroanatomical distribution of neurofibrillary tangles in the medial temporal lobe structures (Braak stages III-V) (Braak and Braak, 1991) is closely related to the presence of impaired consolidation and storage of information (Guillozet et al., 2003). In more advanced AD, as the degeneration and the deposition of neurofibrillary tangles spread to neocortical areas (Braak stages V-VI), other non-memory cognitive domains (e.g., executive function, visuospatial abilities or language) are affected (Nelson et al., 2012).

Atypical AD cases, however, do not clearly fit into the Braak staging scheme and may not be conform to the above-mentioned medial temporal lobe dysfunction paradigm. In particular, a hippocampal-sparing neuropathological pattern has been described in addition to the typical AD and the limbic-predominant subtypes (Murray et al., 2011). Consequently atypical AD cases cannot be early detected by episodic memory tests. Thus, free and cued word list recall tests may be poorly useful in the diagnostic algorithm of such cases (see also the paragraph 3.2.4).

In conclusion, we can consider Phase 1 as *Fully Achieved*. Adequate scientific evidence is indeed available supporting the fact that a) medial temporal lobe structures are affected early in AD and b) the extent of cognitive impairment parallels the severity of cortical neurofibrillary tangles pathology.

3.2 Phase 2 – Clinical assay development for clinical disease

The aim of phase 2 is to define the ability of the biomarker to discriminate patients from controls. The primary aim of this phase is to estimate true positive rate (TPR) and false positive rate (FPR) or receiver operating characteristic (ROC) curve for the assay and to assess its ability to distinguish subjects with and without the disease.

Several studies provide evidence of good/excellent values in the discrimination of AD dementia patients from healthy control subjects for many tasks used to assess episodic memory impairments (Welsh et al., 1991, 1992; Buschke et al., 1997; Ivanoiu et al., 2005; Vogel et al., 2007; Ricci et al., 2012; Sotaniemi et al., 2012; Wolfsgruber et al., 2014; Lemos et al., 2014; Delgado et al., 2016). This is particularly true for the delayed recall score of free and cued word list tasks.

The delayed recall score of the RAVLT task (i.e., learning a 15-word list presented in 5 trials) outperformed (sensitivity 100% and specificity 98.3-100%) the immediate recall (i.e., the sum of learning scores of trials 1-5) in discriminating AD dementia from normal individuals both in the Italian and the Australian samples of Ricci et al.'s study (Ricci et al., 2012). Similarly, a large Finnish study (Sotaniemi et al., 2012) on 171 AD dementia patients and 315 cognitively normal individuals using the memory task of the CERAD (i.e., learning a 10-word list presented in 3 trials), a short neuropsychological battery developed in the US, showed that the delayed recall is the most efficient subtest able to discriminate patients from controls (sensitivity 94% and specificity 85%). This result was further confirmed by Wolfsgruber et al. (2014) who reported excellent cross-sectional diagnostic accuracy for both delayed (AUC 0.943) and immediate recall (AUC 0.936).

The use of controlled learning conditions and of cues in the recall phase also demonstrated very high diagnostic value for distinguishing AD dementia patients from healthy controls (Buschke et al., 1997; Vogel et al., 2007; Lemos et al., 2014; Ivanoiu et al., 2005; Delgado et al., 2016). Lemos et al. (2014) showed that both the total immediate and delayed free recall measures of the FCSRT task had excellent accuracy for AD discrimination (Lemos et al., 2014). Direct comparison of these two measures provided evidence for a better sensitivity of the total free delayed recall (96%) compared to the total free immediate recall (94%) (Lemos et al., 2014). The latter measures showed

however a higher specificity value (99% vs. 97%) (Lemos et al., 2014). **In addition, Delgado et al. (2016) showed that both the “word” and the “picture” versions of the FCSRT have high accuracy (AUC >0.9) in distinguish mild AD from healthy subjects. Although visual cues were easier to recall than verbal cues, resulting in greater scores particularly at the total recall measure, “word” and “picture” performances were well correlated each other (Delgado et al., 2016).**

Direct comparisons between free and cued recall tasks, however, do not provide definitive evidence for a better discrimination validity of one class of tasks with respect to the other. In particular, Vogel et al. (2007) showed high sensitivity (88.6-91.4%) and specificity (96.4%) for the immediate and delayed cued subtests of the Grober-Buschke paradigm, i.e. values comparable with the immediate recall of the ADAS-cog (sensitivity 88.6% and sensitivity 96.4%) (Vogel et al., 2007). The highest sensitivity (100%) was reported for the delayed recall of ADAS-cog (Vogel et al., 2007). Ivanoiu et al. (2005) reported similar findings, showing that the total delayed recall of the cued paradigm RI-48 and the delayed recall of CERAD word list are both sensitive measures to discriminate AD from healthy subjects. In contrast, the savings measure is clearly less sensitive than both delayed recall and immediate recall taken in isolation (Ivanoiu et al., 2005).

In conclusion, sufficient scientific evidence is available to consider Phase 2 main aim as *Fully Achieved*. No data support at present the superiority of a single free or cued recall word list task in the ability to discriminate AD dementia patients from control subjects. Good/excellent discriminating value emerged for all the delayed recall subtests.

3.2.1 Phase 2, Secondary aim 1

The first secondary aim of phase 2 is to optimize procedures for performing the assay and to assess the reproducibility of the assay within and between laboratories.

Limitations are present in current literature on the procedures for the administration of free and cued word list tasks, whose reliability is mandatory to ensure reproducibility across different centers

and experimenters. Despite the fact that all the neuropsychological measures object of this review are available in standardized versions for the reference populations, different administration procedures and formats exist, preventing the direct comparability of results among samples.

Different formats of the same task may result in dissimilar performances (Crane et al., 2012). In particular, some tests seem to be more prone to this effect (e.g., the two RAVLT word lists used in ADNI neuropsychological battery are not equivalent to each other and list 2 is systematically harder than list 1, while the three versions of the ADAS-Cog word list are much more similar to each other) (Crane et al., 2012).

Several versions of the Grober-Buschke paradigm have been developed for the clinical use (Buschke, 1984; Grober and Buschke, 1987; Petersen et al., 1994; Buschke et al., 1997; Ivnik et al., 1997; Grober and Kawas, 1997; Tounsi et al., 1999; O'Connell and Tuokko, 2002; Lekeu et al., 2003; Sarazin et al., 2007; Dierckx et al., 2009; Auriacombe et al., 2010; Frasson et al., 2011; Lemos et al., 2014; Dion et al., 2015). They differ in the number of the to-be-learned items (i.e., 12 or 16), in the physical characteristics of the stimuli (e.g., line drawings, printed or orally spoken words) and in the testing procedures. In addition, the many versions differ in the number of learning trials (i.e., 3 or 6 trials) and in the inclusion of a yes–no recognition test.

The paradigm construction (e.g., presence of low or high number of to-be-remembered items) may considerably influence patient performance, preventing the direct comparison among case series. **Very few are however the attempts to compare the psychometric properties of different word list recall task versions (see for example Delgado et al., 2016).** To date, neither sufficient evidence is available nor has direct effort been employed to systemically compare different standardized versions. For these reasons, this sub-aim can be considered as only *Partly Achieved*.

3.2.2 Phase 2, Secondary aim 2

Secondary aim 2 of Phase 2 is to determine the relationship between biomarker tissue measurements made on tissue (phase 1) and the biomarker measurements made on the noninvasive clinical specimen (phase 2).

This is not applicable to neuropsychological measures. However, since performances on free and cued recall word list tasks may represent indirect measures of neurodegeneration of the medial temporal lobe structures, as reported above (see Phase 1), a predictable correlation between the topographical distribution of AD pathologic changes and the pattern of cognitive impairment appears to exist.

As example, the FCSRT scores have been recently shown to correlate with hippocampal atrophy (Sarazin et al., 2010) and with grey matter loss of medial temporal lobe (Rami et al., 2012; Koric et al., 2013). Moreover, impaired FCSRT performance can be correlated with the presence of AD pathology as shown by changes of CSF Abeta values (Rami et al., 2012; Wagner et al., 2012; Xie et al., 2014), even at a prodromal stage (Derby et al., 2013).

Since no data addressing this specific issue are available, and only indirect evidence is present, we evaluate this sub-aim as *Not Applicable*.

3.2.3 Phase 2, Secondary aim 3

Secondary aim 3 of phase 2 is to assess factors associated with biomarker status or level in control subjects. If such factors affect the biomarker, thresholds for test positivity may need to be defined separately for target sub-populations.

As the majority of cognitive measures, many factors are associated with performances on delayed recall tasks in cognitively healthy subjects. Age, education and gender have a relevant impact on the cognitive performances in neuropsychological testing. Adjusted data of standardized versions account for such variables in the determination of the threshold for test positivity. For this reason, raw scores should be corrected for all the confounding demographic features according to the normative data of the reference population in order to control their weight on the performance.

Thus, normative data are crucial both in clinical settings, to obtain single individual corrected scores, and in research studies aimed at addressing accuracy of test performance. The availability of good normative data, possibly uncontaminated by undiagnosed cognitively impaired control subjects, is thus a necessary prerequisite in order to consider the task as sufficiently reliable.

An important issue to consider when interpreting cognitive test results is the base rate of low performance in the normative sample. For example, 60.6 % of the normative sample of the CERAD neuropsychological assessment battery obtained one or more scores at or below the 10th percentile (Mistridis et al., 2015).

In this context, many tasks included in this review have normative data for many different languages and countries. Efforts for providing normative data for the Grober-Buschke paradigm are more recent, and normative data for FCSRT are available for English-, Spanish-, French-, Italian- and Danish-speaking populations (Ivnik et al., 1997; Grober et al., 1998; Vogel et al., 2007; Frasson et al., 2011; Mokri et al., 2013; Dion et al., 2015).

It must be underlined that the concept of cognitive reserve includes not only the educational level but also to lifestyle and occupation habits (Xu et al., 2015). These factors have significant direct and indirect effects on performances at delayed recall tasks and should therefore be considered in the evaluation and diagnosis of memory complaining subjects (Lojo-Seoane et al., 2014). For example, total recall of FCSRT is more sensitive to cognitive change in MCI subjects at low rather than at high levels of baseline cognition (Mura et al., 2014). The ceiling effect of this task suggests that it should be used with caution to assess memory changes in subjects with high premorbid cognitive performance (Mura et al., 2014).

The weight of demographic variables on the performances at the delayed recall tasks has not been definitely evaluated in all the reference subpopulations. Moreover, apart from educational level, neuropsychological task performances are not controlled for other possible factors influencing cognitive performance at baseline, such as socio-economic status and other factors possibly

affecting cognitive reserve. For these reasons, this sub-aim is to consider as *Partly Achieved*. It is expected that ongoing studies will fill this gap within a few years.

3.2.4 Phase 2, Secondary aim 4

Secondary aim 4 of phase 2 is to assess factors associated with biomarker status or level in diseased subjects—in particular, disease characteristics. The relationship between several disease characteristics (e.g., age at onset, genetic factors, comorbidities, atypical presentation) and episodic memory performances have been partly explored in AD dementia patients.

Memory tasks are not uniformly impaired in early and late onset AD cases (Van der Flier et al., 2011). Multiple studies indicated that early onset of AD is associated with lower performance on different memory tasks (Kalpouzos et al., 2005; Kaiser et al., 2012). **This impairment is strictly intersected with the ApoE carrier condition (Van der Flier et al., 2011). Early-onset AD carriers of the ApoE ϵ 4 allele had worse results for memory tasks than did non-carriers (Van der Flier et al., 2011) and the homozygous ϵ 4/ ϵ 4 genotype is frequently seen in patients with amnesic presentations than in those with non-memory phenotypes (Van der Flier et al., 2011).** ApoE ϵ 4 allele condition is also a well-established risk factor for late-onset AD affecting the age of onset (Shi et al., 2014). **Although the above-mentioned evidence, it is still not clear the weight of ApoE ϵ 4 allele carrier condition is the cognitive decline of AD. Results of a meta-analysis on 1700 patients (Allan and Ebmeier, 2011) showed indeed that different ApoE alleles do not modify the speed of clinical progression of dementia in terms of cognitive decline or mortality.**

Finally, an atypical topographical localization of the neuronal injury in the course of AD pathology (as in case of posterior cortical atrophy, logopenic variant of primary progressive aphasia or frontal variant of AD) may result in a relative sparing of memory function in the early phase of disease.

We can thus consider this sub-aim as only *Partly Achieved*, since further studies are required. Notably, investigations on additional genetic risk factors and on the influence of comorbidities, such as those with diabetes or sleep apnea, are still lacking.

3.3 Phase 3 – Prospective longitudinal repository studies

The primary aims of Phase 3 consist in evaluating the capacity of the biomarker to detect pre-clinical disease (Primary aim 1) and in defining criteria for a positive biomarker test in preparation for phase 4 (Primary aim 2).

Different studies explored the capability of delayed recall tasks to detect the earliest stages of the AD pathology, i.e., MCI who will progress to AD dementia. One point in need of an adequate consideration is the criterion for positivity to be used for neuropsychological tests. In the case of MCI, the threshold for positivity is the same as for AD dementia, namely a neuropsychological performance between one and two standard deviations below an age- and education-matched control population (Petersen, 2004). A reduced performance on memory tasks compared with that of the reference population is required to qualify a cognitive complaint as “objective”. Thus, this is a mandatory prerequisite for defining the condition of MCI and distinguishing it from subjective memory complaints (Petersen, 2004; DSM-V, APA 2013).

The cut-off between normal and pathological findings to be applied on individual basis is crucial for the clinical use of any cognitive marker. Standardization procedure and normative data according to the reference population (see Phase 2, Secondary Aim 3) allow us to clearly define cognitive performance abnormality at the individual level. A challenge, however, could be the determination of specific cut-off points for the prodromal AD phase accounting for possible confounding variables (see Gainotti et al., 2014 for a critical review).

Predictive values of both learning and retention measures of free word list tasks have been widely investigated and have been consistently shown to predict progressions from MCI to AD in both community-based and memory clinic samples (Estevez-Gonzalez et al., 2003; Ivanoiu et al.,

2005; Anchisi et al., 2005; Tabert et al., 2006; Fleisher et al., 2007; Chang et al., 2010; Landau et al., 2010; Kim et al., 2010, Gallagher et al., 2010).

A meta-analytic study on the cognitive impairments in prodromal AD subjects (Bäckman et al., 2005) reported values of immediate and delayed recall subtests for predicting progression to dementia, with a better performance of delayed recall. In agreement with this, Chang et al. showed good predictive values both for immediate and delayed recall of RAVLT, with better predictive power when both measures are impaired (Chang et al., 2010). Literature findings are, however, scanty and heterogeneous in methodology, failing to confirm or disprove the superiority of one measure above the other. As example, Landau et al. (2010) reported very high sensitivity (93%), high specificity (88%) and overall accuracy (90%) for the immediate recall of RAVLT, but in this study this was the only RAVLT measure taken into account. In line with these authors (Landau et al., 2010), Ivanoiu et al. (2005) reported also very high sensitivity for the immediate recall (100%) of the CERAD 10-word list, even better than the savings (90%) and delayed recall (86%) measures. Both immediate and delayed recall scores of the CERAD were good predictors of incident AD dementia ($AUC = 0.835-0.838$) in elderly primary care patients (Wolfsgruber et al., 2014).

Comparable diagnostic accuracy and predictive values have been reported for the delayed recall subtest of other word list paradigms, namely CVLT (Anchisi et al., 2005), Seoul verbal learning test (SVLT) (Kim et al., 2010), ADAS-Cog 10-word list (Fleisher et al., 2007), WMS-III (Espinosa et al., 2013), delayed word recall (DWR) task (Gallagher et al., 2010) and for the percent savings from immediate to delayed recall of the Selective Reminding Test (Tabert et al., 2006). In addition to delayed recall, a more stringent clinical classification (i.e., probable MCI according to Petersen's criteria), the presence of multiple cognitive impairments and of at least one APOE $\epsilon 4$ allele resulted in an increased risk of progression (Espinosa et al., 2013).

Cued paradigms showed also very high predictive value for progression to AD in large studies on MCI subjects (Grober et al., 2000, 2010; Ivanoiu et al., 2005; Sarazin et al., 2007; Derby et al., 2013; Mura et al., 2014; Belleville et al., 2014). The cued delayed recall of the RI48-test showed a

higher accuracy (sensitivity = 100% and specificity = 77%) compared to the CERAD 10-word list (Ivanoiu et al., 2005). Out of 251 MCI subjects tested by Sarazin et al. (2007) at baseline and followed at 6-month intervals for up to 3 years, 59 subjects progressed to AD dementia. Among different neuropsychological tasks, the immediate total recall and the index of cueing of FCSRT had the best accuracy (AUC = 94% and 93%; sensitivity = 79.7% and 78%; specificity = 89.9% and 84.9%, respectively) for the diagnosis of AD in the prodromal phase. Additional results have been reported in a prospective longitudinal study on 122 single-domain or multiple-domain amnesic MCI recruited from memory clinics and followed up to 8 years (Belleville et al., 2014). The strongest predictors were free immediate recall of FCSRT together with the delayed recall of a story (i.e., MEMO-TEXT). The best model, yielding high sensitivity (88%), specificity (87.1%), and very high positive (81.8%) and negative (91.7%) predictive values included both delayed recall tasks and non-memory measures, supporting the utility of combining different neuropsychological tasks to obtain higher accuracy (Belleville et al., 2014).

In conclusion, the literature review is generally supporting good predictive values for both immediate and delayed recall measures. Some comparative studies among different tasks (e.g., one free recall task vs. another; free vs. cued paradigms; different memory paradigms with each other) are available, with sparse and contradictory findings. Word list recall tests with a built-in semantic self-cueing option (i.e., CVLT) has been shown to be more sensitive to subtle episodic memory impairments compared to simple unrelated word list learning (Beck et al., 2012). While it is well established that scores of different free word list tasks (e.g., RAVLT and CVLT) highly correlate with each other in cognitively normal individuals (Crossen and Wiens, 1994), unquestionable similarities or relevant differences have not been clearly shown in neurodegenerative patients. Literature findings are inconsistent, with some authors stressed the lack of critical differences between free or cued tasks in the prediction of MCI conversion to AD dementia (Ivanoiu et al., 2005), whereas others supported better operating characteristics for the cued paradigm compared to free word list learning and/or logical memory measures (Derby et al., 2013; Wagner et al., 2012). In

addition, better predictive values have been suggested for the combination of delayed recall tasks with other cognitive measures (i.e., Mini Mental Status Examination and Mental Control subtest of the Wechsler memory scale) (Pozueta et al., 2011; Tierney et al., 1996). **Recently, a composite episodic memory score proved in MCI better performance compared to individual memory scores to improve the power to predict progression to AD (Marra et al., 2015).** Unfortunately, no direct comparison between a given format of a recall task vs. another one in the same population is available.

There is at the moment insufficient evidence for the superiority of one measure compared to the others, as well as of cued paradigms compared to traditional measures of free recall. Further studies are needed to address the impact of different cut-off points for delayed recall tests in MCI. As reported by the Bruscoli and Lovestone's (2004) in a systematic review of conversion studies, the more stringent the measures of memory impairment, the better the prediction of conversion. Therefore, it seems essential to explore this issue in future research on large longitudinal studies. We can thus consider primary aims of Phase 3 studies as only *Partly Achieved*. Some evidence is in fact available but many methodological limitations (e.g., lack of sufficient direct comparisons among the different task measures) prevent to draw final conclusions.

3.3.1 Phase 3, Secondary aim 1

Secondary aim 1 of phase 3 is to explore the impact of covariates on the discriminatory abilities of the biomarker before clinical diagnosis.

This issue has already been discussed above in relation to age, education and gender (Phase 2 Secondary Aim 3). Several additional factors that may affect the effectiveness of covariates on test scores in discriminating at risk patients from control subjects are occupational history, socio-economic status as well as contextual factors (time of the day, metabolic status). These have not been systematically investigated.

Since no data addressing the above-mentioned factors are available, we can thus consider secondary aim 1 of Phase 3 studies as only *Partly Achieved*.

3.3.2 Phase 3, Secondary aim 2

The secondary aim 2 of phase 3 is to compare markers with a view to selecting those that are most promising.

As the performance on memory test is better considered as a gateway to the application of biomarkers, rather than a biomarker *per se*, a direct comparison between memory test scores and AD biomarkers in terms of diagnostic contribution is not considered as appropriate.

This sub-aim is thus evaluated as *Not Applicable*.

3.3.3 Phase 3, Secondary aim 3

The aim is to develop algorithms for positivity based on combination of markers.

No specific diagnostic algorithm for positivity based on combination of delayed recall tests and AD biomarkers has been yet proposed. The relevance of including neuropsychological markers in the diagnostic algorithm is incontestable, as cognitive measures contribute to define the population of interest (e.g., amnesic or non-amnesic MCI subjects; AD or non-AD dementia patients) (see introduction for the concept of objective memory performance as a “gatekeeper” to biomarker assessment).

Several studies however tested the combination of memory measures (not directly used for the classification of the patient) with structural and/or functional imaging and CSF markers in early or prodromal AD cases. A number of papers, for example, reported predictive values of cognitive measures combined with MRI (Visser et al., 2002; Geroldi et al., 2006; Devanand et al., 2007; Heister et al., 2011; Ewers et al., 2012; Richard et al., 2013; Eckerström et al., 2013), CSF (Richard et al., 2013; Eckerström et al., 2013) or FDG-PET imaging (Chételat et al., 2005; Anchisi et al., 2005; Landau et al., 2010).

Richard et al. (2013) found no incremental value of adding up MRI and CSF biomarkers to the outcome of a memory test (i.e., immediate recall of RAVLT) in the early differential diagnosis of patients with MCI (average follow-up of 39 months). This study simulates well the clinical reality by first establishing diagnostic accuracy for the neuropsychology task, and then quantifying the performance of MRI and CSF biomarkers. All three diagnostic instruments (i.e., delayed recall test, MRI imaging and CSF values) substantially contributed to the differentiation of stable MCI from converters to AD (AUCs ~ 0.65), while neither MRI nor CSF significantly improved diagnostic accuracy compared to the RAVLT administration alone. Different data have been reported by other authors. For instance, Eckerström et al. (2013) reported that, while delayed recall of RAVLT is the best individual predictor of dementia compared to CSF values and hippocampal volume measurement, an incremental diagnostic value is provided by the combination of these diagnostic instruments. The three measures combined (i.e., RAVLT, CSF and MRI) were, indeed, the most successful combination for predicting dementia (AUC = 0.96) (Eckerström et al., 2013). Additionally, Heister et al. (2011) supported the utility of combining different tools to increase the diagnostic value. These authors reported an increased risk of AD conversion in subjects with joint presence of any two risk factors among immediate recall of RAVLT, MRI and CSF (Heister et al., 2011). The combination of greater learning impairment and increased atrophy was associated with the highest risk (namely, 85% of patients with both risk factors converted to AD within 3 years) (Heister et al., 2011). In further support of the additional value of combining different diagnostic measures are the results of the study of Landau et al. (2010) in which the authors showed that MCI patients with abnormal results on both FDG-PET imaging and RAVLT task were 11.7 times more likely to progress to AD than subjects who had normal results on both measures.

In conclusion, the literature review is generally supporting improved predictive values for the combination of delayed recall tasks with biological and imaging biomarkers, with a particular impact of the latter. The evidence is, however, scanty and mainly focused on the combination of RAVLT performance with imaging or CSF biomarkers. No conclusion can, thus, be drawn on the

superiority of a specific combination of one neuropsychological tool and topographical/biological biomarkers relative to another combination (e.g., RAVLT or CVLT or cued paradigms combined to CSF measures or MRI and PET imaging), since the available evidence is not sufficient and not yet sufficiently replicated with adequately powered methods.

In particular, the lack of evidence on which is the delayed recall task to prefer forces us to consider this sub-aim of Phase 3 as only *Partly Achieved*.

3.3.4 Phase 3, Secondary aim 4

This aims to determine a biomarker testing interval for phase 4 if repeated testing is of interest.

There are several problematic issues in the application of the concept to this area. First, the delayed recall tests tend to go to floor very fast, and are thus of limited usefulness to study disease progression (Locascio et al., 1995). Repeated testing is of interest only in the very early and prodromal disease stages, when subtle cognitive changes may be detected only with repeated measurements. Second, in the very early and prodromal stage an important issue is availability of parallel forms to account for learning effects. For example, the same performance at follow-up may indicate a lack of practice effects, i.e. a minimal (initial) disorder (Zehnder et al., 2007). Unfortunately, this is still a largely neglected area, with some exceptions (i.e., RBANS).

On the basis of these considerations, this sub-aim is at present to consider as *Not Achieved*.

3.4 Phase 4 – Prospective diagnostic studies

The primary aim of Phase 4 is to determine the operating characteristics of the biomarker-based test in a relevant population by determining the detection rate and the false referral rate. Studies at this stage involve testing clinical populations and leading to diagnosis and treatment.

The main aim of Phase 4 is thus still *Not Achieved*.

3.4.1 Phase 4, Secondary aim 1

Secondary aim 1 of phase 4 assesses the characteristics of disease detected by the biomarker test—in particular, with regard to the potential benefit incurred by early detection.

Only preliminary evidence is available in the case of AD. In general, the scientific and medical community agrees upon the fact that earlier diagnosis may be advantageous for patients and caregivers, particularly for optimized medical management, future planning, access to services, risk reduction, and delaying nursing home placement (Banerjee et al., 2007; Banerjee and Wittenberg 2009; Relkin, 2000).

This sub-aim is thus evaluated as *Not Achieved*.

3.4.2 Phase 4, Secondary aim 2

The secondary aim 2 of phase 4 assesses the practical feasibility of implementing the diagnostic program and compliance of test-positive subjects with work-up and treatment recommendations.

This sub-aim has not been assessed by any specific study, and is thus evaluated as *Not Achieved*.

3.4.3 Phase 4, Secondary aim 3

The aim is to make preliminary assessments of the effects of biomarker testing on costs and mortality associated with the disease.

This aspect has not been specifically assessed. This aim is thus evaluated as *Not Achieved*.

3.4.4 Phase 4, Secondary aim 4

This aim is to monitor disease occurring clinically but not detected by the biomarker testing protocol.

This sub-aim has not been specifically assessed and thus evaluated as *Not Achieved*.

3.5 Phase 5 – Disease control studies

This final phase addresses whether using biomarkers for early diagnosis reduces the burden of Alzheimer's disease in the general population.

Given the lack of effective pharmacological and/or non pharmacological interventions for AD, population-level screening for dementia may not be acceptable to either the general public or health care professionals (Martin et al., 2015). Stigma, financial implications, impacts on employment and driving, as well as on person's lifestyle and attitude to health, combined with the knowledge of poor prognosis and few treatment options, play a crucial role in decision-making about population screening.

Thus, primary aim and sub-aims of Phase 5 have not been specifically assessed and should be evaluated as *Not Achieved*.

4. Conclusions

A comprehensive neuropsychological evaluation of cognitive functions is required in all patients with subjective complaint not otherwise explained in order to identify the early stages of neurodegeneration. The choice of the task to assess the different cognitive domains is critical, in view of reaching sufficient sensitivity and specificity to detect early subjects at risk of developing AD dementia (i.e., prodromal AD cases). The psychometric properties of the test (e.g., sensitivity to cognitive change, specificity for neurodegeneration, floor- and ceiling- effects) have an important impact. The neuropsychological measures used to determine the MCI condition should be considered as the entry point to appropriate diagnostic algorithm for AD or non-AD neurocognitive disorders, as they make it possible to delineate the target population (in research settings) or the patients (in clinical settings), which may be candidate to undergo other invasive (i.e., CSF), costly (i.e., MRI) or both invasive and costly (i.e., PET imaging) biomarker assessments. The proper use of single neuropsychological tasks and, even better, the development of a combination of tasks are particularly urgent in the diagnosis of neurodegenerative dementias. The objective confirmation of

long-term memory impairment, which is the earliest and most prominent feature of typical AD dementia (Salmon, 2012), is in need of further specific investigation. Noteworthy, no recommendation concerning the choice of the memory task to be used to assess this function in MCI subjects is currently available. The use of cued memory paradigms is suggested (Dubois et al., 2010, 2014) but no specific guidelines have been provided so far regarding the format that should be preferred. During the last 15 years, cognitive markers have been extensively used, in spite of the absence of clear-cut guidelines, according to the availability of the tasks in the specific research or clinical setting, mainly in academic memory clinics.

This review is part of a wider effort to outline a roadmap for the correct application of AD biomarkers in research and clinical settings. The whole framework is borrowed from drug development, and its specific translation for the oncology field (Pepe et al., 2001). We have revised the framework in order to adapt it to the current context and to the needs of the field of neurodegenerative dementias with a focus on the maturity of free and cued word list measures. Impairment on this class of tasks is considered as the core of neuropsychological feature of typical AD dementia, consistent with the neuroanatomic distribution of histopathological abnormalities in medial temporal regions already present in the mildest stages of AD (Braak and Braak, 1991). Some authors previously reviewed neuropsychological predictors of conversion from MCI to AD dementia (see for examples Gainotti et al., 2014; Modrego, 2006; Li et al., 2015). Here, we included only studies providing evidence for diagnostic accuracy in AD or for the prediction of progression from MCI to AD, with the main aim to unveil grey areas in need of further confirmation.

The limitations of this review reflect the inhomogeneous and non-systematic use of neuropsychological testing in research and clinical settings. Despite the variability among studies in design, statistical approach, follow-up length, and cognitive tests used, a general pattern seems to emerge. The majority of studies on AD dementia patients reported delayed recall as the most sensitive measure. Higher delayed recall scores on auditory verbal learning tasks seem to be 'protective' for the progression to AD dementia, as shown by a recent meta-analysis and review of

cohort studies assessing risk factors in MCI subjects (Li et al., 2015). The literature is also in favor of good predictive values for the progression from MCI to AD dementia not only for delayed but also for immediate recall measures. Although this literature review provided no definitive support for the relative advantage of cued recall tasks in comparison with tests measuring free recall in screening of patients suspected for AD dementia, cued paradigms seem to offer better operating characteristics compared to free word list learning (Wagner et al., 2012). There is, however, clearly the need for systematic studies directly comparing delayed recall from free or cued tasks. A clear-cut conclusion on the preferable use of a specific measure vs. another cannot be drawn at the moment, this remains a critical topic that should be addressed by future research aimed at comparing several neuropsychological tests (and scores) within the memory domain.

We focused on the assessment of delayed recall since its impairment is a key feature of typical AD dementia. Nevertheless, this will not be an accurate neuropsychological measure in atypical AD cases (e.g., posterior cortical atrophy, logopenic variant of primary progressive aphasia, and frontal variant of AD). Therefore, this measure may not always be the best choice to orient towards a screening with second-level biomarkers in suspicion of an AD condition. A combination of memory and non-memory neuropsychological measures should thus be preferred as the gateway to the assessment of second-level invasive and costly biomarkers. Accordingly, different neuropsychological measures taken together may provide a better accuracy than the measurement of single cognitive domains (see for examples Tierney et al., 1996; Tabert et al., 2006; Mitchell, 2009a-b; Dierckx et al., 2009; Belleville et al., 2014). Further research is required to determine the best cognitive algorithm (namely, a combination of different memory tasks or memory and non-memory tests) that is expected to have a higher predictive diagnostic value, because it reflects the different components of the neurobiological dysfunctions occurring in AD pathology.

This review has several limitations. Notwithstanding our efforts to be as inclusive as possible, the literature search was not a formal systematic review. Moreover, the framework adapted by Pepe et al. (2001) to the oncology field had to be adapted to the current context, i.e. to provide a reliable

and biomarker-based diagnosis in people with cognitive complaints. This required adaptation of the screening phases into a diagnostic perspective, which is conceptually different and has relevant limitations compared to screening studies. Future development of the field may allow to expand the currently developed framework and to fully assimilate it to the oncology and drug-development framework.

5. Acknowledgements

The Geneva Task Force for the Roadmap of Alzheimer's Biomarkers includes the participants to a workshop held in Geneva on December 8-9, 2014. The P.I. of the Geneva Roadmap effort is Giovanni B Frisoni, with Bengt Winblad and Clifford R Jack Jr as co-PIs. The task force includes experts in biomarker development from the oncology community; experts on diagnostic AD biomarkers from Switzerland and Europe; representatives of pertinent scientific societies (Federation of European Societies of Neuropsychology - FENS, European Neurological Society of Neuroradiology - ENSNR, International Foundation of Clinical Chemistry and Laboratory Medicine - IFCCLM, European Association of Nuclear Medicine - EANM, and Swiss Federation of Clinical Neuro Societies - SFCNS); representatives of patient advocates, bioethicists and regulatory agencies, and early career researchers. The Geneva Task Force has been endorsed by the EADC – European Alzheimer's Disease Consortium.

The workshop was funded thanks to a competitive grant by the Swiss National Science Foundation and unrestricted grants from: Alzheimer Forum Switzerland, Association pour la Recherche sur Alzheimer, Genève; Piramal, Eli Lilly & Company, General Electric, Guerbet, TEVA Pharma; Academie Suisse de Sciences Médicales, Vifor Pharma Switzerland, Novartis, Siemens, and IXICO. The Alzheimer's Association hosted the first follow-up meeting of the initiative at the 2015 AAIC congress in Washington. We acknowledge the help from Margherita Mauri and Daria Gennaro (IRCCS Fatebenefratelli, Brescia, Italy) and Agnese Picco (Università di Genova, Genova, Italy) who took care of the logistics of the workshop.

The following scientific societies took part to the Geneva Workshop for the Roadmap of Alzheimer's Biomarkers on December 8-9 2014. Flavio Nobili was delegate from the European Association of Nuclear Medicine (EANM) Neuroimaging Committee. Kaj Blennow was delegate and Chair of the International Federation of Clinical Chemistry and Laboratory Medicine Working Group for CSF proteins (IFCC WG-CSF). Frederik Barkhof was delegate from the European Society of Neuroradiology (ESNR). Stefano Cappa was delegate and Chair of the Federation of European Societies of Neuropsychology (FENS). Urs Mosimann was delegate from the Swiss Federation of Clinical Neuro Societies (SFCNS). The content of this paper represents the opinion of the individual authors and is not necessarily endorsed by the scientific societies which took part to the Geneva Workshop for the Roadmap of Alzheimer's Biomarkers.

Disclosure statement

Chiara Cerami was funded by Fondazione Eli-Lilly grant 2011 "Imaging of neuroinflammation and neurodegeneration in prodromal and presymptomatic Alzheimer's disease phases". Marina Boccardi received a research grant from Piramal. Andreas U. Monsch and Stefano F. Cappa have nothing to declare.

References

- Abdulrab K, Heun R. Subjective Memory Impairment. A review of its definitions indicates the need for a comprehensive set of standardised and validated criteria. *Eur Psychiatry* 2008;23(5):321–30.
- Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH. 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. *Alzheimers Dement* 2011;7(3):270–79.

- Allan CL, Ebmeier KP. The influence of ApoE4 on clinical progression of dementia: a meta-analysis. *Int J Geriatr Psychiatry* 2011;26(5):520–6.
- Anchisi D, Borroni B, Franceschi M, Kerrouche N, Kalbe E, Beuthien-Beumann B, Cappa S, Lenz O, Ludecke S, Marcone A, Mielke R, Ortelli P, Padovani A, Pelati O, Pupi A, Scarpini E, Weisenbach S, Herholz K, Salmon E, Holthoff V, Sorbi S, Fazio F, Perani D. Heterogeneity of brain glucose metabolism in mild cognitive impairment and clinical progression to Alzheimer disease. *Arch Neurol* 2005;62(11):1728–33.
- APA, American Psychiatric Association. Diagnostic and statistical manual of mental disorders IV-TR. 4th ed. text revised. Washington, DC; 2000.
- Auriacombe S, Helmer C, Amieva H, Berr C, Dubois B, Dartigues JF. Validity of the free and cued selective reminding test in predicting dementia: the 3C study. *Neurology*. 2010;74(22):1760–77.
- Bäckman L, Jones S, Berger AK, Laukka EJ, Small BJ. Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology* 2005;19(4):520–31.
- Banerjee S, Willis R, Matthews D, Contell F, Chan J, Murray J. Improving the quality of care for mild to moderate dementia: an evaluation of the Croydon Memory Service Model. *Int J Geriatr Psychiatry* 2007;22(8):782–8.
- Banerjee S, Wittenberg R. Clinical and cost effectiveness of services for early diagnosis and intervention in dementia. *Int J Geriatr Psychiatry* 2009;24(7):748–54.
- Barbeau E, Didic M, Tramon E, Felician O, Joubert S, Sontheimer A, Ceccaldi M, Poncet M. Evaluation of visual recognition memory in MCI patients. *Neurology* 2004;62:1317–22.
- Barbeau EJ, Ranjeva JP, Didic M, Confort-Gouny S, Felician O, Soulier E, Cozzone PJ, Ceccaldi M, Poncet M. Profile of memory impairment and gray matter loss in amnesic mild cognitive impairment. *Neuropsychologia* 2008;46:1009–19.
- Beck IR, Gagneux-Zurbriggen A, Berres M, Taylor KI, Monsch AU. Comparison of verbal episodic memory measures: consortium to establish a registry for Alzheimer's disease-

Neuropsychological Assessment Battery (CERAD-NAB) versus California Verbal Learning Test (CVLT). *Arch Clin Neuropsychol* 2012;27(5):510–9.

- Belleville S, Gauthier S, Lepage E, Kergoat MJ, Gilbert B. Predicting decline in mild cognitive impairment: a prospective cognitive study. *Neuropsychology* 2014;28(4):643–52.
- Bennett DA, Wilson RS, Schneider JA, Evans DA, Beckett LA, Aggarwal NT, Barnes LL, Fox JH, Bach J. Natural history of mild cognitive impairment in older persons. *Neurology* 2002;59(2):198–205.
- Berrios GE, Markova IS, Giralá N. Functional memory complaints: hypochondria and disorganization. In *Memory disorders in psychiatric practice*. Cambridge University Press; 2000. pp. 384–99.
- Braak H, Braak E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol* 1991;82(4):239–59.
- Brandt J, Benedict RHB. Hopkins Verbal Learning Test—Revised. Professional manual. Lutz, FL: Psychological Assessment Resources, Inc; 2001.
- Bruscoli M, Lovestone S. Is MCI really just early dementia? A systematic review of conversion studies. *Int Psychogeriatr* 2004;16(2):129–40.
- Buschke H. Cued recall in amnesia. *J Clin Neuropsychol* 1984;6(4):433–40.
- Buschke H, Sliwinski MJ, Kuslansky G, Lipton RB. Diagnosis of early dementia by the Double Memory Test: encoding specificity improves diagnostic sensitivity and specificity. *Neurology* 1997;48(4):989–97.
- Chang YL, Bondi MW, Fennema-Notestine C, McEvoy LK, Hagler DJ Jr, Jacobson MW, Dale AM; Alzheimer's Disease Neuroimaging Initiative. Brain substrates of learning and retention in mild cognitive impairment diagnosis and progression to Alzheimer's disease. *Neuropsychologia* 2010;48(5):1237–47.
- Chételat G, Eustache F, Viader F, De La Sayette V, Pélerin A, Mézenge F, Hannequin D, Dupuy B, Baron JC, Desgranges B. FDG-PET measurement is more accurate than

neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. *Neurocase* 2005;11(1):14–25.

- Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, Jones RN, Mukherjee S, Curtis SM, Harvey D, Weiner M, Mungas D; Alzheimer's Disease Neuroimaging Initiative. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging Behav* 2012;6(4):502–16.
- Crossen JR, Wiens AN. Comparison of the Auditory-Verbal Learning Test (AVLT) and California Verbal Learning Test (CVLT) in a sample of normal subjects. *J Clin Exp Neuropsychol* 1994;16(2):190–4.
- **Delgado C, Muñoz-Neira C, Soto A, Martínez M, Henríquez F, Flores P, Slachevsky A. Comparison of the Psychometric Properties of the "Word" and "Picture" Versions of the Free and Cued Selective Reminding Test in a Spanish-Speaking Cohort of Patients with Mild Alzheimer's Disease and Cognitively Healthy Controls. Arch Clin Neuropsychol. 2016 Jan 11. pii: acv107. [Epub ahead of print]**
- Delis DC, Kramer JH, Kaplan E, Ober BA. The California Verbal Learning Test. New York: The Psychological Corporation, San Antonio, TX; 1987.
- Delis DC, Kramer JH, Kaplan E, Ober BA. The California Verbal Learning Test (2nd ed.). New York: The Psychological Corporation, San Antonio, TX; 2000.
- Derby CA, Burns LC, Wang C, Katz MJ, Zimmerman ME, L'italien G, Guo Z, Berman RM, Lipton RB. Screening for predementia AD: time-dependent operating characteristics of episodic memory tests. *Neurology* 2013;80(14):1307–14.
- Devanand DP, Pradhaban G, Liu X, Khandji A, De Santi S, Segal S, Rusinek H, Pelton GH, Honig LS, Mayeux R, Stern Y, Tabert MH, de Leon MJ. Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease. *Neurology*. 2007;68(11):828–36.

- Dierckx E, Engelborghs S, De Raedt R, Van Buggenhout M, De Deyn PP, Verté D, Ponjaert-Kristoffersen I. Verbal cued recall as a predictor of conversion to Alzheimer's disease in Mild Cognitive Impairment. *Int J Geriatr Psychiatry*. 2009;24(10):1094–100.
- Dickerson BC, Eichenbaum H. The episodic memory system: neurocircuitry and disorders. *Neuropsychopharmacology* 2010;35(1):86–104.
- Dion M, Potvin O, Belleville S, Ferland G, Renaud M, Bherer L, Joubert S, Vallet GT, Simard M, Rouleau I, Lecomte S, Macoir J, Hudon C. Normative data for the Rappel libre/Rappel indicé à 16 items (16-item Free and Cued Recall) in the elderly Quebec-French population. *Clin Neuropsychol* 2015;28(Suppl 1):S1–19.
- Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol* 2007;6(8):734–46.
- Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, Jicha GA, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Sarazin M, de Souza LC, Stern Y, Visser PJ, Scheltens P. Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol* 2010;9(11):1118–27.
- Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R, Cappa S, Crutch S, Engelborghs S, Frisoni GB, Fox NC, Galasko D, Habert MO, Jicha GA, Nordberg A, Pasquier F, Rabinovici G, Robert P, Rowe C, Salloway S, Sarazin M, Epelbaum S, de Souza LC, Vellas B, Visser PJ, Schneider L, Stern Y, Scheltens P, Cummings JL. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol* 2014;13(6):614–29.

- Eckerström C, Olsson E, Bjerke M, Malmgren H, Edman A, Wallin A, Nordlund A. A combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts conversion from mild cognitive impairment to dementia. *J Alzheimers Dis* 2013;36(3):421–31.
- Eichenbaum H, Cohen NJ. *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford University Press; 2001.
- Elias MF, Beiser A, Wolf PA, Au R, White RF, D’Agostino RB. The preclinical phase of Alzheimer disease. A 22-year prospective study of the Framingham Cohort. *Arch Neurol* 2000;57:808–13.
- Espinosa A, Alegret M, Valero S, Vinyes-Junqué G, Hernández I, Mauleón A, Rosende-Roca M, Ruiz A, López O, Tárraga L, Boada M. A longitudinal follow-up of 550 mild cognitive impairment patients: evidence for large conversion to dementia rates and detection of major risk factors involved. *J Alzheimers Dis* 2013;34(3):769–80.
- Estevez-Gonzalez A, Kulisevsky J, Boltes A, Otermin P, Garcia-Sanchez C. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer’s disease: comparison with mild cognitive impairment and normal aging. *Int J Geriatr Psychiatr* 2003;18:1021–8.
- Ewers M, Walsh C, Trojanowski JQ, Shaw LM, Petersen RC, Jack CR Jr, Feldman HH, Bokde AL, Alexander GE, Scheltens P, Vellas B, Dubois B, Weiner M, Hampel H; North American Alzheimer's Disease Neuroimaging Initiative (ADNI). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiol Aging* 2012;33(7):1203–14.
- Fleisher AS, Sowell BB, Taylor C, Gamst AC, Petersen RC, Thal LJ; Alzheimer's Disease Cooperative Study. Clinical predictors of progression to Alzheimer disease in amnesic mild cognitive impairment. *Neurology* 2007;68(19):1588–95.

- Frasson P, Ghiretti R, Catricalà E, Pomati S, Marcone A, Parisi L, Rossini PM, Cappa SF, Mariani C, Vanacore N, Clerici F. Free and Cued Selective Reminding Test: an Italian normative study. *Neurol Sci* 2011;32(6):1057–62.
- Frisoni GB, Perani D, Bastianello S, Bernardi G, Porteri C, Boccardi M, Cappa SF, Trabucchi M, Padovani A. A roadmap to the use of biomarkers for the diagnosis of Alzheimer's disease in clinical practice: the Italian inter-societal consensus. *Neurobiol Aging* 2015, this issue (Submitted, September 8, 2015; NBA 15-700).
- Gainotti G, Quaranta D, Vita MG, Marra C. Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease. *J Alzheimers Dis* 2014;38(3):481–95.
- Gallagher D, Mhaolain AN, Coen R, Walsh C, Kilroy D, Belinski K, Bruce I, Coakley D, Walsh JB, Cunningham C, Lawlor BA. Detecting prodromal Alzheimer's disease in mild cognitive impairment: utility of the CAMCOG and other neuropsychological predictors. *Int J Geriatr Psychiatry* 2010;25(12):1280–7.
- Geroldi C, Rossi R, Calvagna C, Testa C, Bresciani L, Binetti G, Zanetti O, Frisoni GB. Medial temporal atrophy but not memory deficit predicts progression to dementia in patients with mild cognitive impairment. *J Neurol Neurosurg Psychiatry* 2006;77(11):1219–22.
- Grober E, Buschke H. Genuine memory deficits in dementia. *Dev Neuropsychol* 1987;3:13–36.
- Grober E, Kawas C. Learning and retention in preclinical and early Alzheimer's disease. *Psychol Aging* 1997;12(1):183–8.
- Grober E, Lipton RB, Katz M, Sliwinski M. Demographic influences on free and cued selective reminding performance in older persons. *J Clin Exp Neuropsychol* 1998;20(2):221–6.
- Grober E, Lipton RB, Hall CB, Crystal H. Memory impairment on free and cued selective reminding predicts dementia. *Neurology* 2000;54(4):827–32.
- Grober E, Sanders AE, Hall C, Lipton RB. Free and cued selective reminding identifies very mild dementia in primary care. *Alzheimer Dis Assoc Disord* 2010;24(3):284–90.

- Guillozet AL, Weintraub S, Mash DC, Mesulam MM. Neurofibrillary tangles, amyloid, and memory in aging and mild cognitive impairment. *Arch Neurol* 2003;60:729–36.
- Hartley T, Bird CM, Chan D, Cipolotti L, Husain M, Varga-Khadem F, Burgess N. The hippocampus is required for short term topographical memory in humans. *Hippocampus* 2007;17:34–48.
- Heister D, Brewer JB, Magda S, Blennow K, McEvoy LK; Alzheimer's Disease Neuroimaging Initiative. Predicting MCI outcome with clinically available MRI and CSF biomarkers. *Neurology* 2011;77(17):1619–28.
- Hornberger M, Piguet O, Graham AJ, Nestor PJ, Hodges JR. How preserved is episodic memory in behavioral variant frontotemporal dementia? *Neurology* 2010;74(6):472–9.
- Ivanoiu A, Adam S, Van der Linden M, Salmon E, Juillerat AC, Mulligan R, Seron X. Memory evaluation with a new cued recall test in patients with mild cognitive impairment and Alzheimer's disease. *J Neurol* 2005;252(1):47–55.
- Ivnik RJ, Smith GE, Lucas JA, Tangalos EG, Kokmen E, Petersen RC. Free and cued selective reminding test: MOANS norms. *J Clin Exp Neuropsychol* 1997;19:676–91.
- Jack CR Jr, Lowe VJ, Senjem ML, Weigand SD, Kemp BJ, Shiung MM, Knopman DS, Boeve BF, Klunk WE, Mathis CA, Petersen RC. 11C PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnesic mild cognitive impairment. *Brain* 2008;131(Pt 3):665–80.
- Jack CR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7(3):257–62.
- Jain KK. Biomarkers of Neurological Disorders. In *Applications of Biotechnology in Neurology*. Humana Press; 2012. pp. 49–153.

- Jessen F, Amariglio RE, van Boxtel M, Breteler M, Ceccaldi M, Chtelat G, Dubois B, Dufouil C, Ellis KA, van der Flier WM, Glodzik L, van Harten AC, de Leon MJ, McHugh P, Mielke MM, Molinuevo JL, Mosconi L, Osorio RS, Perrotin A, Petersen RC, Rabin LA, Rami L, Reisberg B, Rentz DM, Sachdev PS, de la Sayette V, Saykin AJ, Scheltens P, Shulman MB, Slavin MJ, Sperling RA, Stewart R, Uspenskaya O, Vellas B, Visser PJ, Wagner M; Subjective Cognitive Decline Initiative (SCD-I) Working Group. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimers Dement* 2014;10(6):844–52.
- Kaiser NC, Melrose RJ, Liu C, Sultzer DL, Jimenez E, Su M, Monserratt L, Mendez MF. Neuropsychological and neuroimaging markers in early versus late-onset Alzheimer's disease. *Am J Alzheimers Dis Other Dement* 2012;27(7):520–9.
- Kalpouzos G, Eustache F, de la Sayette V, Viader F, Chtelat G, Desgranges B. Working memory and FDG-PET dissociate early and late onset Alzheimer disease patients. *J Neurol* 2005;252(5):548–58.
- Kim SH, Seo SW, Yoon DS, Chin J, Lee BH, Cheong HK, Han SH, Na DL. Comparison of neuropsychological and FDG-PET findings between early- versus late-onset mild cognitive impairment: A five-year longitudinal study. *Dement Geriatr Cogn Disord* 2010;29(3):213–23.
- Koric L, Ranjeva JP, Felician O, Guye M, de Anna F, Soulier E, Didic M, Ceccaldi M. Cued recall measure predicts the progression of gray matter atrophy in patients with amnesic mild cognitive impairment. *Dement Geriatr Cogn Disord* 2013;36(3-4):197–210.
- Landau SM, Harvey D, Madison CM, Reiman EM, Foster NL, Aisen PS, Petersen RC, Shaw LM, Trojanowski JQ, Jack CR Jr, Weiner MW, Jagust WJ; Alzheimer's Disease Neuroimaging Initiative. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology* 2010;75(3):230–8.

- Lekeu F, Van der Linden M, Chicherio C, Collette F, Degueldre C, Franck G, Moonen G, Salmon E. Brain correlates of performance in a free/cued recall task with semantic encoding in Alzheimer disease. *Alzheimer Dis Assoc Disord* 2003;17(1):35–45.
- Lemos R, Simões MR, Santiago B, Santana I. The free and cued selective reminding test: Validation for mild cognitive impairment and Alzheimer's disease. *J Neuropsychol* 2015;9(2):242–57.
- Li JQ, Tan L, Wang HF, Tan MS, Tan L, Xu W, Zhao QF, Wang J, Jiang T, Yu JT. Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies. *J Neurol Neurosurg Psychiatry* 2015 May 22. pii: jnnp-2014-310095. doi: 10.1136/jnnp-2014-310095. [Epub ahead of print]
- Locascio JJ, Growdon JH, Corkin S. Cognitive test performance in detecting, staging, and tracking Alzheimer's disease. *Arch Neurol* 1995;52(11):1087–99.
- Lojo-Seoane C, Facal D, Guàrdia-Olmos J, Juncos-Rabadán O. Structural model for estimating the influence of cognitive reserve on cognitive performance in adults with subjective memory complaints. *Arch Clin Neuropsychol* 2014;29(3):245–55.
- Lowndes G, Savage G. Early detection of memory impairment in Alzheimer's disease: a neurocognitive perspective on assessment. *Neuropsychol Rev* 2007;17:193–202.
- **Marra C, Gainotti G, Fadda L, Perri R, Lacidogna G, Scaricamazza E, Piccininni C, Quaranta D. Usefulness of an Integrated Analysis of Different Memory Tasks to Predict the Progression from Mild Cognitive Impairment to Alzheimer's Disease: The Episodic Memory Score (EMS). *J Alzheimers Dis.* 2015 Nov 27. [Epub ahead of print]**
- Martin S, Kelly S, Khan A, Cullum S, Denning T, Rait G, Fox C, Katona C, Cosco T, Brayne C, LaFortune L. Attitudes and preferences towards screening for dementia: a systematic review of the literature. *BMC Geriatr* 2015 Jun 16;15:66.

- Maruta C, Guerreiro M, de Mendonça A, Hort J, Scheltens P. The use of neuropsychological tests across Europe: the need for a consensus in the use of assessment tools for dementia. *Eur J Neurol* 2011;18(2):279–85.
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984;34:939–44.
- McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7(3):263–9.
- Meltzer JA, Constable RT. Activation of human hippocampal formation reflects success in both encoding and cued recall of paired associates. *Neuroimage* 2005;24(2):384–97.
- Mistridis P, Egli SC, Iverson GL, Berres M, Willmes K, Welsh-Bohmer KA, Monsch AU. Considering the base rates of low performance in cognitively healthy older adults improves the accuracy to identify neurocognitive impairment with the Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery (CERAD-NAB). *Eur Arch Psychiatry Clin Neurosci* 2015;265(5):407–17.
- Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J Psychiatr Res* 2009a;43(4):411–31.
- Mitchell J, Arnold R, Dawson K, Nestor PJ, Hodges JR. Outcome in subgroups of mild cognitive impairment (MCI) is highly predictable using a simple algorithm. *J Neurol* 2009b;256(9):1500–9.
- Modrego PJ. Predictors of conversion to dementia of probable Alzheimer type in patients with mild cognitive impairment. *Curr Alzheimer Res* 2009;3(2):161–70.

- Mohs RC, Rosen WG, Davis KL. The Alzheimer's Disease Assessment Scale: An instrument for assessing treatment efficacy. *Psychopharmacol Bull* 1983;19(3):448–50.
- Mokri H, Avila-Funes JA, Meillon C, Gutiérrez Robledo LM, Amieva H. Normative data for the Mini-Mental State Examination, the Free and Cued Selective Reminding Test and the Isaacs Set Test for an older adult Mexican population: the Coyoacán cohort study. *Clin Neuropsychol* 2013;27(6):1004–18.
- Morris JC, Mohs RC, Rogers H, Fillenbaum G, Heyman A. Consortium to establish a registry for Alzheimer's disease (CERAD) clinical and neuropsychological assessment of Alzheimer's disease. *Psychopharmacol Bull* 1988;24(4):641–52.
- Mura T, Proust-Lima C, Jacqmin-Gadda H, Akbaraly TN, Touchon J, Dubois B, Berr C. Measuring cognitive change in subjects with prodromal Alzheimer's disease. *J Neurol Neurosurg Psychiatry* 2014;85(4):363–70.
- **Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol*. 2011 Sep; 10(9):785–796.**
- Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 2005;53(4):695–9.
- Nelson PT, Alafuzoff I, Bigio EH, Bouras C, Braak H, Cairns NJ, Castellani RJ, Crain BJ, Davies P, Del Tredici K, Duyckaerts C, Frosch MP, Haroutunian V, Hof PR, Hulette CM, Hyman BT, Iwatsubo T, Jellinger KA, Jicha GA, Kövari E, Kukull WA, Leverenz JB, Love S, Mackenzie IR, Mann DM, Masliah E, McKee AC, Montine TJ, Morris JC, Schneider JA, Sonnen JA, Thal DR, Trojanowski JQ, Troncoso JC, Wisniewski T, Woltjer RL, Beach TG. Correlation of Alzheimer disease neuropathologic changes with cognitive status: a review of the literature. *J Neuropathol Exp Neurol* 2012;71(5):362–81.

- O'Connell ME, Tuokko H. The 12-item Buschke memory test: appropriate for use across levels of impairment. *Appl Neuropsychol* 2002;9(4):226–33.
- Parra MA, Abrahams S, Fabi K, Logie R, Luzzi S, Della Sala S. Short-term memory binding deficits in Alzheimer's disease. *Brain* 2009;132(Pt 4):1057–66.
- Parra MA, Abrahams S, Logie RH, Méndez LG, Lopera F, Della Sala S. Visual short-term memory binding deficits in familial Alzheimer's disease. *Brain* 2010;133(9):2702–13.
- Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 2001;93(14):1054–61.
- Petersen RC, Smith GE, Ivnik RJ, Kokmen E, Tangalos EG. Memory function in very early Alzheimer's disease. *Neurology* 1994;44(5):867–72.
- Petersen RC. Mild cognitive impairment as a diagnostic entity. *J Intern Med* 2004;256(3):183–94.
- Pozueta A, Rodríguez-Rodríguez E, Vazquez-Higuera JL, Mateo I, Sánchez-Juan P, González-Perez S, Berciano J, Combarros O. Detection of early Alzheimer's disease in MCI patients by the combination of MMSE and an episodic memory test. *BMC Neurol* 2011;11:78.
- Rami L, Solé-Padullés C, Fortea J, Bosch B, Lladó A, Antonell A, Olives J, Castellví M, Bartres-Faz D, Sánchez-Valle R, Molinuevo JL. Applying the new research diagnostic criteria: MRI findings and neuropsychological correlations of prodromal AD. *Int J Geriatr Psychiatry* 2012;27(2):127–34.
- Randolph C, Gold JM, Kozora E, Cullum M, Hermann BP, Wyler AR. Estimating memory function: disparity of Wechsler Memory Scale – Revised and California Verbal Learning Test indices in clinical and normal samples. *The Clinical Neuropsychologist* 1994;8:99–108.
- Relkin N. Screening and early diagnosis of dementia. *Am J Manag Care* 2000;6(22 Suppl):S1111–1118; discussion S1119–1124.

- Rey A. L'examen psychologique dans les cas d'encéphalopathie traumatique. Archives de Psychologie 1941;28:21.
- Ricci M, Graef S, Blundo C, Miller LA. Using the Rey Auditory Verbal Learning Test (RAVLT) to differentiate alzheimer's dementia and behavioural variant fronto-temporal dementia. Clin Neuropsychol 2012;26(6):926–41.
- Richard E, Schmand BA, Eikelenboom P, Van Gool WA; Alzheimer's Disease Neuroimaging Initiative. MRI and cerebrospinal fluid biomarkers for predicting progression to Alzheimer's disease in patients with mild cognitive impairment: a diagnostic accuracy study. BMJ Open. 2013 Jun 20; 3(6). pii: e002541. doi: 10.1136/bmjopen-2012-002541.
- Rowe CC, Ellis KA, Rimajova M, Bourgeat P, Pike KE, Jones G, Fripp J, Tochon-Danguy H, Morandau L, O'Keefe G, Price R, Raniga P, Robins P, Acosta O, Lenzo N, Szoëke C, Salvado O, Head R, Martins R, Masters CL, Ames D, Villemagne VL. Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging. Neurobiol Aging 2010;31(8):1275–83.
- Salmon DP. Neuropsychological features of mild cognitive impairment and preclinical Alzheimer's disease. Curr Top Behav Neurosci 2012;10:187–212.
- Sarazin M, Berr C, De Rotrou J, Fabrigoule C, Pasquier F, Legrain S, Michel B, Puel M, Volteau M, Touchon J, Verny M, Dubois B. Amnesic syndrome of the medial temporal type identifies prodromal AD: a longitudinal study. Neurology 2007;69:1859–67.
- Sarazin M, Chauviré V, Gerardin E, Colliot O, Kinkingnéhun S, de Souza LC, Hugonot-Diener L, Garnero L, Lehericy S, Chupin M, Dubois B. The amnesic syndrome of hippocampal type in Alzheimer's disease: an MRI study. J Alzheimers Dis 2010;22(1):285–94.
- Shi J, Han P, Kuniyoshi SM. Cognitive impairment in neurological diseases: lessons from apolipoprotein E. J Alzheimers Dis 2014;38(1):1–9.

- Small BJ, Fratiglioni L, Viitanen M, Winblad B, Backman L. The course of cognitive impairment in preclinical Alzheimer disease. Three- and 6-year follow-up of a population-based sample. *Arch Neurol* 2000;57:839–44.
- Sotaniemi M, Pulliainen V, Hokkanen L, Pirttilä T, Hallikainen I, Soininen H, Hänninen T. CERAD-neuropsychological battery in screening mild Alzheimer's disease. *Acta Neurol Scand* 2012;125(1):16–23.
- Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR Jr, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7(3):280–92.
- Squire LR, Stark CE, Clark RE. The medial temporal lobe. *Annu Rev Neurosci* 2004;27:279–306.
- Tabert MH, Manly JJ, Liu X, Pelton GH, Rosenblum S, Jacobs M, Zamora D, Goodkind M, Bell K, Stern Y, Devanand DP. Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment. *Arch Gen Psychiatry* 2006;63(8):916–24.
- Tierney MC, Szalai JP, Snow WG, Fisher RH, Nores A, Nadon G, Dunn E, St George-Hyslop PH. Prediction of probable Alzheimer's disease in memory-impaired patients: A prospective longitudinal study. *Neurology* 1996;46(3):661–5.
- Tounsi H, Deweer B, Ergis AM, Van der Linden M, Pillon B, Michon A, Dubois B. Sensitivity to semantic cuing: an index of episodic memory dysfunction in early Alzheimer disease. *Alzheimer Dis Assoc Disord* 1999;13(1):38–46.
- Tulving E. Episodic memory: from mind to brain. *Annu Rev Psychol* 2002;53:1–25.

- **van der Flier WM, Pijnenburg YA, Fox NC, Scheltens P. Early-onset versus late-onset Alzheimer's disease: the case of the missing APOE ε4 allele. Lancet Neurol. 2011 Mar; 10(3): 280–8.**
- Visser PJ, Verhey FR, Hofman PA, Scheltens P, Jolles J. Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. J Neurol Neurosurg Psychiatry 2002;72(4):491–7.
- Vogel A, Mortensen EL, Gade A, Waldemar G. The Category Cued Recall test in very mild Alzheimer's disease: discriminative validity and correlation with semantic memory functions. Eur J Neurol 2007;14(1):102–8.
- Wagner M, Wolf S, Reischies FM, Daerr M, Wolfsgruber S, Jessen F, Popp J, Maier W, Hüll M, Frölich L, Hampel H, Perneczky R, Peters O, Jahn H, Luckhaus C, Gertz HJ, Schröder J, Pantel J, Lewczuk P, Kornhuber J, Wiltfang J. Biomarker validation of a cued recall memory deficit in prodromal Alzheimer disease. Neurology 2012;78:379–86.
- Welsh K, Butters N, Hughes J, Mohs R, Heyman A. Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. Arch Neurol 1991;48:278–81.
- Welsh KA, Butters N, Hughes JP, Mohs RC, Heyman A. Detection and staging of dementia in Alzheimer's disease. Use of the neuropsychological measures developed for the Consortium to Establish a Registry for Alzheimers Disease. Arch Neurol 1992;49:448–52.
- Wolfsgruber S, Jessen F, Wiese B, Stein J, Bickel H, Mösch E, Weyerer S, Werle J, Pentzek M, Fuchs A, Köhler M, Bachmann C, Riedel-Heller SG, Scherer M, Maier W, Wagner M; AgeCoDe Study Group. The CERAD neuropsychological assessment battery total score detects and predicts Alzheimer disease dementia with high diagnostic accuracy. Am J Geriatr Psychiatry 2014;22(10):1017–28.
- Xie J, Gabelle A, Dorey A, Garnier-Crussard A, Perret-Liaudet A, Delphin-Combe F, Bathsavanis A, Dauphinot V, Lehmann S, Mercier B, Desestret V, Rouillet-Solignac I, Vighetto

A, Krolak-Salmon P. Initial memory deficit profiles in patients with a cerebrospinal fluid Alzheimer's disease signature. *J Alzheimers Dis* 2014;41(4):1109–16.

- Xu W, Yu JT, Tan MS, Tan L. Cognitive reserve and Alzheimer's disease. *Mol Neurobiol* 2015; 51(1):187–208.
- Zehnder AE, Bläsi S, Berres M, Spiegel R, Monsch AU. Lack of practice effects on neuropsychological tests as early cognitive markers of Alzheimer disease? *Am J Alzheimers Dis Other Dement* 2007;22(5):416–26.

Table 1. The table reports evaluation terms and assessments of aims and sub-aims of each Phase.

Phase	General Aim	Aim	Progress	Evidence in AD
1 – Preclinical exploratory studies	Identify the rational of the BM, based on pathology	To identify leads for potentially useful biomarkers and prioritize identified leads	Fully achieved	Adequate scientific evidence is available supporting the fact that a) medial temporal lobe structures are early affected in AD and b) the extent of cognitive impairment parallels the severity of cortical neurofibrillary tangles pathology.
2 – Clinical Assay development for Clinical Disease	Define the ability of the BM to discriminate patients from controls	<i>Primary:</i> To estimate True Positive Rate (TPR) and False Positive Rate (FPR) or Receiver Operating Characteristic (ROC) curve for the assay and to assess its ability to distinguish subjects with and without disease	Fully achieved	Good/excellent value for the delayed recall tasks emerged in the discrimination of AD patients from cognitively healthy subjects.
		<i>Secondary 1:</i> To optimize procedures for performing the assay and to assess the reproducibility of the assay within and between laboratories	Partly achieved	Neither sufficient evidence is available nor has direct effort been employed to systemically compare different standardized versions of delayed recall tasks.
		<i>Secondary 2:</i> To determine the relationship between biomarker tissue measurements made on tissue (phase 1) and the biomarker measurements made on the noninvasive clinical specimen (phase 2)	Not applicable	-
		<i>Secondary 3:</i> To assess factors (e.g. sex, age, etc.), associated with biomarker status or level in control subjects. If such factors affect the biomarker, thresholds for test positivity may need to be defined separately for target subpopulations.	Partly achieved	The weight of demographic variables on the performances at the delayed recall tasks has not been definitely evaluated in all the reference subpopulations. Apart from educational level, neuropsychological task performances are not controlled for other possible factors influencing cognitive performance at baseline, such as socio-economic status and other factors possibly affecting cognitive reserve.
		<i>Secondary 4:</i> To assess factors associated with biomarker status or level in diseased subjects—in particular, disease characteristics.	Partly achieved	Neuropsychological performances on delayed memory tasks correlate with age at onset. ApoE ε4 allele presence is not directly associated with cognitive status. Episodic memory performances may not be sensitive to atypical AD. Additional studies assessing factors associated with biomarker status in AD subjects are required.
3 –	Define the ability	<i>Primary 1:</i> To evaluate the capacity of the	Partly	Scientific evidence is generally supporting good predictive

Retrospective Longitudinal Repository Studies	of the BM to detect the disease in its early phase	biomarker to detect the earliest disease stages	achieved	values for delayed recall tasks. Some comparative studies among different tasks are available, with sparse and contradictory findings. There is at the moment insufficient evidence for the superiority of one measure compared to the others, as well as of cued paradigms compared to traditional measures of free recall.
		<i>Primary 2:</i> To define criteria for a biomarker positive test in preparation for phase 4.	Partly achieved	It seems essential to address the impact of different cut-off points for delayed recall tests in future large longitudinal studies on MCI subjects.
		<i>Secondary 1:</i> To explore the impact of covariates on the discriminatory abilities of the biomarker before clinical diagnosis.	Partly achieved	Performances on episodic memory tasks are associated with demographic variables and cognitive reserve. Several factors that may affect the effectiveness of covariates on test scores in discriminating at risk patients from control subjects (e.g., occupational history, socio-economic status and contextual factors such as time of the day or metabolic status) have not been systematically investigated.
		<i>Secondary 2:</i> To compare markers with a view to selecting those that are most promising	Not applicable	-
		<i>Secondary 3:</i> To develop algorithms for positivity based on combinations of markers.	Partly achieved	The literature review is generally supporting improved predictive values for the combination of delayed recall tasks with biological and imaging biomarkers, with a particular impact of the latter. The evidence is, however, scanty and mainly focused on the combination of RAVLT performance with imaging or CSF biomarkers.
		<i>Secondary 4:</i> To determine a biomarker testing interval for phase 4 if repeated testing is of interest.	Not achieved	There are several problematic issues in this area. First, the delayed recall tests tend to go to floor very fast, and are thus of limited usefulness to study disease progression. Repeated testing is of interest only in the very early and prodromal disease stages, when subtle cognitive changes may be detected only with repeated measurements. Second, in the very early and prodromal stage an important issue is availability of parallel forms to account for learning effects.

4 – Prospective Case Finding Studies	Quantify the BM accuracy and usefulness in patients diagnosed and treated based on the BM	<i>Primary:</i> To determine the operating characteristics of the biomarker-based test in a relevant population by determining the detection rate and the false referral rate. Studies at this stage involve testing people and lead to diagnosis and treatment.	Not achieved	-
		<i>Secondary 1:</i> To describe the characteristics of disease detected by the biomarker test—in particular, with regard to the potential benefit incurred by early detection.	Not achieved	-
		<i>Secondary 2:</i> To assess the practical feasibility of implementing the case finding program and compliance of test-positive subjects with work-up and treatment recommendations.	Not achieved	-
		<i>Secondary 3:</i> To make preliminary assessments of the effects of biomarker testing on costs and mortality associated with the disease.	Not achieved	-
		<i>Secondary 4:</i> To monitor disease occurring clinically but not detected by the biomarker testing protocol.	Not achieved	-
5 – Disease Control Studies	Quantify the impact of the BM-based diagnosis on clinically meaningful outcomes and costs	<i>Primary:</i> To estimate the reductions in disease-associated mortality, morbidity, and disability afforded by biomarker testing.	Not achieved	-
		<i>Secondary 1:</i> To obtain information about the costs of biomarker testing and treatment and the cost per life saved or per quality-adjusted life year	Not achieved	-
		<i>Secondary 2:</i> To evaluate compliance with testing and work-up in a diverse range of settings.	Not achieved	-
		<i>Secondary 3:</i> To compare different biomarker testing protocols and/or to compare different approaches to treating test positive subjects in	Not achieved	-

		regard to effects on mortality and costs.		
--	--	---	--	--