Journal of Alzheimer's Disease 46 (2015) 1079–1089
DOI 10.3233/JAD-143015
IOS Press

Novel Statistically-Derived Composite Measures for Assessing the Efficacy of Disease-Modifying Therapies in Prodromal Alzheimer's Disease Trials: An AIBL Study

Samantha C. Burnham[a,*], Nandini Raghavan[b], William Wilson[c], David Baker[d], Michael T. Ropacki[e], Gerald Novak[b], David Ames[f,g], Kathryn Ellis[f], Ralph N. Martins[h,i], Paul Maruff[j], Colin L. Masters[k], Gary Romano[b], Christopher C. Rowe[l,m], Greg Savage[n], S. Lance Macaulay[o], Vaibhav A. Narayan[b], for the Alzheimer's Disease Neuroimaging Initiative[1] and the AIBL Research Group[2]

[a]*CSIRO Digital Productivity Flagship, Floreat, WA, Australia*
[b]*Janssen Research and Development, Raritan, NJ, USA*
[c]*CSIRO Digital Productivity Flagship, North Ryde, NSW, Australia*
[d]*Janssen Research and Development, Titusville, NJ, USA*
[e]*Janssen Research and Development, Fremont, CA, USA*
[f]*Academic Unit for Psychiatry of Old Age, Department of Psychiatry, University of Melbourne, Parkville, VIC, Australia*
[g]*National Ageing Research Institute, Parkville, VIC, Australia*
[h]*Centre of Excellence for Alzheimer's Disease Research & Care, School of Medical Sciences, Edith Cowan University, Joondalup, WA, Australia*
[i]*Sir James McCusker Alzheimer's Disease Research Unit (Hollywood Private Hospital), Perth, WA, Australia*
[j]*Cogstate, Melbourne, VIC, Australia*
[k]*Mental Health Research Institute (MHRI), The University of Melbourne, Parkville, VIC, Australia*
[l]*Department of Nuclear Medicine and Centre for PET, Austin Health, Heidelberg, VIC, Australia*
[m]*Department of Medicine, Austin Health, The University of Melbourne, Heidelberg, VIC, Australia*
[n]*ARC Centre of Excellence in Cognition and its Disorders, and Department of Psychology, Macquarie University, Sydney, NSW, Australia*
[o]*CSIRO Food and Nutrition Flagship, Melbourne, VIC, Australia*

Accepted 3 April 2015

Abstract.
Background: There is a growing consensus that disease-modifying therapies must be given at the prodromal or preclinical stages of Alzheimer's disease (AD) to be effective. A major unmet need is to develop and validate sensitive measures to track disease progression in these populations.

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[2]http://www.aibl.csiro.au/about/aibl-research-team

*Correspondence to: Samantha Burnham, DPF, CSIRO, Private Bag 5, Wembley, WA 6913, Australia. Tel.: +61 0 8 9333 6706; Fax: +61 0 8 9333 6121; E-mail: samantha.burnham@csiro.au.

ISSN 1387-2877/15/$35.00

Objective: To generate novel statistically-derived composites from standard scores, which have increased sensitivity in the assessment of change from baseline in prodromal AD.
Methods: An empirically based method was employed to generate domain specific, global, and cognitive-functional novel composites. The novel composites were compared and contrasted with each other, as well as standard scores for their ability to track change from baseline. The longitudinal characteristics and power to detect decline of the measures were evaluated. Data from participants in the Australian Imaging, Biomarkers and Lifestyle (AIBL) Study characterized as mild cognitively impaired with high neocortical amyloid-β burden were utilized for the study.
Results: The best performing standard scores were CDR Sum-of-Boxes and MMSE. The statistically-derived novel composites performed better than the standard scores from which they were derived. The domain-specific composites generally did not perform as well as the global composites or the cognitive-functional composites.
Conclusion: A systematic method was employed to generate novel statistically-derived composite measures from standard scores. Composites comprised of measures including function and multiple cognitive domains appeared to best capture change from baseline. These composites may be useful to assess progression or lack thereof in prodromal AD. However, the results should be replicated and validated using an independent clinical sample before implementation in a clinical trial.

Keywords: Alzheimer's disease, clinical marker, clinical trial, mild cognitive impairment, prodromal stage

INTRODUCTION

Alzheimer's disease (AD) is considered a global priority. Given aging populations and the increasing incidence of AD, health and economic frameworks are set to be crippled. Thus, there exists an unprecedented challenge to understand and cure this disease.

In this quest, clinical research is moving away from therapeutic trials aimed at countering symptomatic effects and toward therapeutic trials aimed at modifying the underlying disease mechanism. The majority of current trials focus on altering the deposition of extracellular amyloid-β (Aβ) plaques, one of the hallmark histological markers of AD.

To successfully prevent the development of AD, the growing consensus is that disease modifying therapies may need to be given early; most likely at the presymptomatic stage. One of the key challenges faced by such studies is the uncertainty over appropriate endpoints for prodromal (or even preclinical) AD trials. Therefore, a major unmet need is to develop and validate sensitive measures to track disease progression in prodromal and/or preclinical AD populations.

Historically, assessment in AD clinical trials leveraged the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog) [1], which demonstrates utility in monitoring longitudinal cognitive performance in AD dementia cohorts. However, the ADAS-Cog has known insensitivity in early AD populations [2–5], which raises the question of suitability for prodromal trials.

A ceiling effect is manifest when measures are insensitive to differences in the population of interest: i.e., most patients perform at the maximum score and thus, differentiating between those who have modest impairments and those who are functioning normally is not possible. Alternatively, a test may exhibit a floor effect where there is a lack of sensitivity in identifying differences in the population of interest. This occurs because most patients score at the minimum possible score thereby making it impossible to differentiate patients at the diseased end of the spectrum, to show additional longitudinal change (i.e., decline), or to differentiate between patients over time.

Ideally, clinical endpoints would have a dynamic range of performance across the population of interest, hence, not exhibit ceiling or floor effects. Different measures will capture performance most accurately at different points of the disease spectrum. Thus, the combination of multiple measures into a single composite would likely represent a desirable clinical endpoint for a broader range of the disease spectrum. A composite may also provide greater statistical power over evaluating multiple individual tests, and reduce the risk of spurious inferences.

Episodic memory, executive function, and language represent some of the cognitive domains that most often change across the spectrum of AD dementia [6]. However, there is no consensus as to which measures are most efficacious in their ability to track longitudinal progression in prodromal AD or whether creating a composite from those most sensitive measures would improve performance compared to the standard scores. Further, with the use of a composite measure, there is the option to combine multiple scores tapping into various cognitive domains or even to develop mixed cognitive-functional composites: a number of cognitive-functional composites have been considered for endpoints in recent contributions [5, 7, 8].

In this study, domain-specific standard scores, domain-specific composite scores, multi-domain/global composite scores, as well as cognitive-functional composites were compared and contrasted. A systematic approach was adopted to develop novel composite measures of domain-specific, global, and cognitive-functional ability. The ability to detect change from baseline as well as the longitudinal characteristics and power of the measures were evaluated. Data from participants engaged in the Australian Imaging, Biomarkers and Lifestyle (AIBL) Study with mild cognitive impairment (MCI) and evidence of high neocortical Aβ burden were utilized for the study. In an attempt to determine the likely inter-study comparability of standard scores and thus their resultant composite scores a rank comparison of the overlapping standard scores available both in AIBL and the Alzheimer's Disease Neuroimaging Initiative (ADNI) was also completed.

The analyses were restricted to prodromal AD (MCI) participants with high neocortical Aβ burden as it is likely that measures to optimally monitor progression will be specific to the population and severity of disease under consideration. Therefore, preclinical AD participants were not included. Mild cognitively impaired participants with low neocortical Aβ burden were also not included for three primary reasons: a) clinical trials for prodromal AD are now predominantly designed with inclusion criteria that incorporates evidence of a positive biomarker (amyloid) finding; b) in the AIBL MCI population, neocortical Aβ burden has a bimodal distribution suggesting that high and low Aβ burden groups represent differing populations; and c) in the AIBL MCI population, longitudinal cognitive decline was more pronounced in the high versus low Aβ burden groups.

MATERIALS AND METHODS

The Australian Imaging, Biomarkers and Lifestyle (AIBL) study

Detailed information on the study design and enrolment procedures have been reported elsewhere [9]. The AIBL Study is a prospective longitudinal study of aging, integrating data from neuroimaging, biomarkers, lifestyle, clinical, and neuropsychological analysis. Eligible volunteers aged over 60 years and fluent in English were classified into three groups: 1) individuals meeting NINCDS-ADRDA criteria for AD [10]; 2) individuals meeting criteria for MCI [11, 12]; and 3) cognitively healthy individuals (healthy controls; HC). The institutional ethics committees of Austin Health, St Vincent's Health, Hollywood Private Hospital, and Edith Cowan University approved the AIBL study, and all volunteers gave written informed consent before participating.

Alzheimer's Disease Neuroimaging Initiative (ADNI)

For validation purposes, data was obtained from the ADNI [13] database. Details on the study design, enrolment procedures, and sample collection are given in the Supplementary Material.

Neuropsychological evaluation

All participants underwent extensive neuropsychological testing as described previously [9]. Briefly, the tests that comprised the AIBL clinical and neuropsychological battery were selected to cover the main domains of cognition affected by AD and other dementias, and are all internationally recognized as having good evidence of their reliability and validity. The full neuropsychological battery comprised: the Clinical Dementia Rating (CDR), Mini-Mental State Examination (MMSE) [14], Clock Drawing Test, California Verbal Learning Test – Second edition (CVLT-II) [15], Logical Memory (LM) I and II (WMS-III; Story A only) [16–18], D-KEFS verbal fluency [19], 30-item Boston Naming Test (BNT) [20], the Stroop task (Victoria version) [17], the Rey Complex Figure Test (RCFT) [21], Digit Span and Digit Symbol-Coding subtests of the Wechsler Adult Intelligence Scale – Third edition (WAIS–III) [22], and the Wechsler Test of Adult Reading [23].

The standard scores from the tests measured in the AIBL battery were corrected for age, gender, years of education, premorbid IQ (Full Scale Intelligence Quotient), and depression symptoms (Geriatric Depression Scale). Multiple Linear Regressions were used to derive the norms from a within-AIBL population comprising 592 HC study participants, who remained healthy over a 3-year period [24]. These within-study norms represent longitudinally robust norms that are able to mitigate increased variability and reduced means that could impact on the ability to monitor progression [25].

PET imaging

A total of 55 AIBL MCI subjects underwent baseline ^{11}C Pittsburg Compound-B (PiB) positron emission

tomography (PET) imaging. The PiB imaging methodology is detailed elsewhere [26]. Spatially normalized PiB-PET images were scaled using the cerebellar grey matter as reference region for the generation of standardized uptake value ratios (SUVR). Neocortical Aβ burden was expressed as the average SUVR of the mean of frontal, superior parietal, lateral temporal, lateral occipital, and anterior and posterior cingulate regions. A SUVR cutoff of 1.5, determined through a cluster analysis of HC individuals, was used to classify participants as belonging to a high or low neocortical Aβ burden group.

Statistical analysis

Of the 55 AIBL MCI subjects who underwent baseline PiB-PET imaging, 4 were excluded due to no baseline neuropsychological test scores being available and a further 14 were excluded as they were classified as having low neocortical Aβ burden. A similar method to that reported by Raghavan et al. [5] was applied to the remaining MCI subpopulation of 37 individuals. Specifically, these subjects would have met NIA-AA criteria for MCI in AD [6, 11, 12, 27], based on a classification of high neocortical Aβ burden and significant cognitive impairments in one or more domains; presence of a neuronal injury biomarker, though a component of the NIA-AA criteria, was not used to define this population. The standard scores (corrected for age, gender, years of education, premorbid IQ, and depressive symptomology, and normed to Z-Scores based on within-study HC participants, who remained healthy over a 3-year period [24]) were assessed to find those with the highest magnitude of change from baseline. Magnitude of change was examined using F-statistics from linear mixed effect (LME) models over the 36 month follow-up period as well as F-statistics from general least squares (GLS) models of 18 month and 36 month changes from baseline. LME was used primarily as it incorporates the full longitudinal dataset, but as LME assumes linearity throughout the time course, GLS analyses for the 18 and 36 month follow-up periods were also examined to circumvent this assumption. Analyses were subject to 100 times 10-fold cross-validation to compute 95% confidence intervals for all F-statistics. The longitudinal trajectories for each of the standard scores were assessed visually using boxplots and quantified using Student's *t*-tests.

Domain-specific composites were generated for verbal episodic memory, visual episodic memory, executive function, and language. This was achieved by taking the Z-Scores (normed to within-study HC participants who remained healthy over a 3-year period) of the sum of the standard scores associated with the cognitive domain of interest which exhibited the greatest magnitudes of F-statistics. Two statistically-derived global composites were also generated by taking Z-Scores of the sum of the standard scores, irrespective of their cognitive domain, with the greatest magnitudes of F-statistics.

The methodology described above for determining the magnitude of change from baseline and assessing longitudinal trajectories for the existing standard scores was also applied to the derived novel composites.

Power and sample size calculations [28] were performed for standard scores and the novel composites to assess the comparative efficacy of the measures as endpoints for clinical trials. These analyses were based upon 3-year LME models with a random intercept and random slope, power calculations were for a two-arm parallel design clinical trial with a hypothesized treatment effect of 25%. All calculations were subject to 100 times 10-fold cross-validation to compute 95% confidence intervals.

Validation

To determine the reproducibility of our results, a validation subset of participants from ADNI was obtained, comprising 60 individuals who were classified as MCI and had a positive Aβ scan at baseline. Eight overlapping standard scores were compared. The ADNI standard scores were normalised to within ADNI study HC norms generated from 170 HC participants who remained healthy over a three year period. Norms were generated for age, gender, years of education and depression symptoms in the same manner as carried out for the AIBL dataset.

The magnitudes of change from baseline of the standard scores available both in the AIBL and the ADNI clinical and neuropsychological batteries were compared in a rank comparison test in an attempt to establish inter-study reproducibility. Rankings were achieved by comparing the magnitude of F-statistics from LME models of baseline, 18 & 36 month follow-up for each cohort.

RESULTS

Demographics

Due to the strict criteria, only 37 individuals were evaluated in this study. The mean age of the subpopulation was 77 years and a high proportion of the

participants (75%) were carriers of the apolipoprotein E ε4 (*APOE* ε4) allele; see Table 1.

Change over time metrics for the standard scores

Twenty-seven standard scores from the AIBL clinical and neuropsychological test battery were initially evaluated: CDR sum of boxes (CDRsb); MMSE; Clock Drawing Test (Clock); measures of learning, short delayed free recall, long delayed free recall, recognition hits, false positives and recognition discrimination d' from CVLT-II ($CVLT_{Lrn}$, $CVLT_{SDFR}$, $CVLT_{LDFR}$, $CVLT_{Hits}$, $CVLT_{FP}$, $CVLT_{d}$); LMI and LMII; Letter Fluency (LetFl), Category Fluency (animals, boys names; CatFl), Category Switch Total ($CatSw_{Tot}$) and Category Switch Accuracy ($CatSw_{Acc}$) from D-KEFS verbal fluency; BNT; Dots, Words, Colours and Colours/Dots Interference scores from the Stroop test ($Stroop_{Dt}$, $Stroop_{Wd}$, $Stroop_{Cl}$, $Stroop_{Int}$); copy, three & 30 min delayed recall, time-to-copy and recognition from RCFT (RCFT, $RCFT_{3DR}$, $RCFT_{30DR}$, $RCFT_{tm}$, $RCFT_{rec}$); and WAIS-III Digit Symbol-Coding (DSC) and Digit Span (DSp).

Table 1
Demographics

Descriptor	Metric	Baseline
Number of Participants	n	37
Age (years)	Mean (sd)	77.14 (6.23)
Female	n (%)	20 (54.05)
<9 years of education	n (%)	4 (10.81)
9–12 years of education	n (%)	15 (40.54)
13–15 years of education	n (%)	9 (24.32)
>15 years of education	n (%)	9 (24.32)
ApoE ε4 carriers	n (%)	28 (75.68)
Premorbid IQ	Mean (sd)	111.24 (6.57)
Geriatric Depression Scale	Mean (sd)	1.90 (1.18)
Standardized uptake value ratio	Mean (sd)	2.21 (0.40)

The magnitudes of change from baseline, represented by the 100 × 10-fold cross-validated F-statistics computed from LME models over the 36 month follow-up period for each of the standard scores are given by the white bars in Fig. 1. It can be seen that the global measures (CDRsb, MMSE, and Clock) and measures of episodic memory ($CVLT_{d}$, $CVLT_{FP}$, and LMII) have the highest magnitudes of F-statistics. The same pattern is also reflected in the 18 and 36 month GLS models (Supplementary Figure 1).

Longitudinal performance of the standard scores

Thirty-six month trajectories for nine of the standard scores are given in Fig. 2A–J. These incorporate the five standard scores with the highest F-statistics (CDRsb, MMSE, Clock, LMII, and $CVLT_{FP}$) as well as domain-specific tests for Verbal Episodic Memory ($CVLT_{LDFR}$), Visual Episodic Memory ($RCFT_{30DR}$), Executive Function ($Stroop_{Int}$), and Language (CatFl). Supplementary Figure 2A–J details the trajectories of these measures for all disease stages.

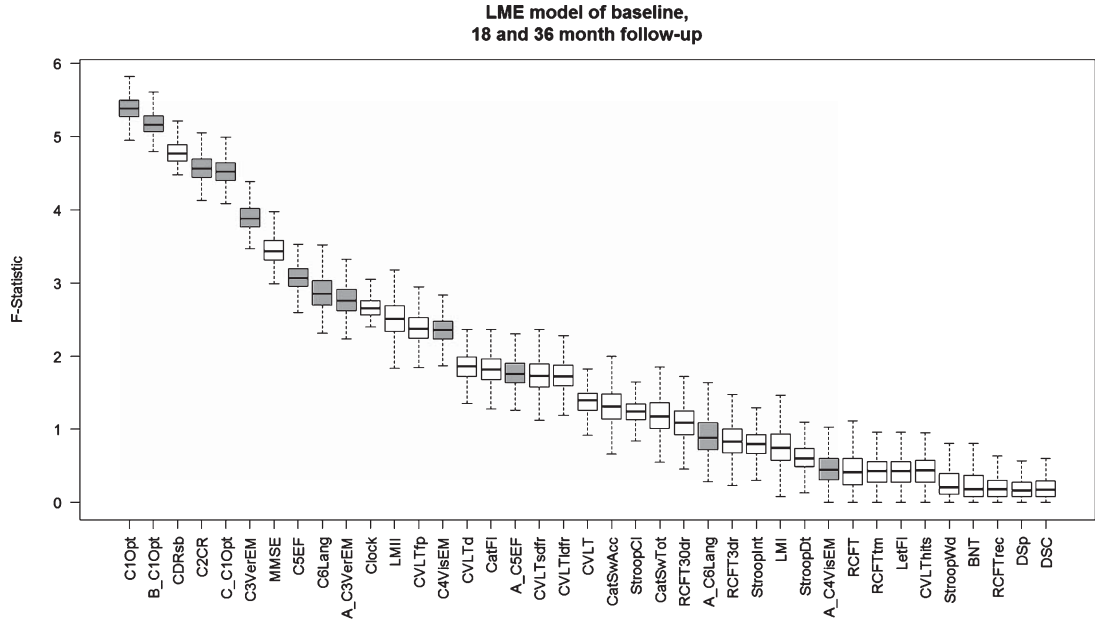

Fig. 1. Comparison of change with time statistics for the standard scores and novel composites. Boxplots of 100 × 10-fold cross-validated F-statistics for LME models are given. White boxes represent standard scores and grey boxes represent novel composites.

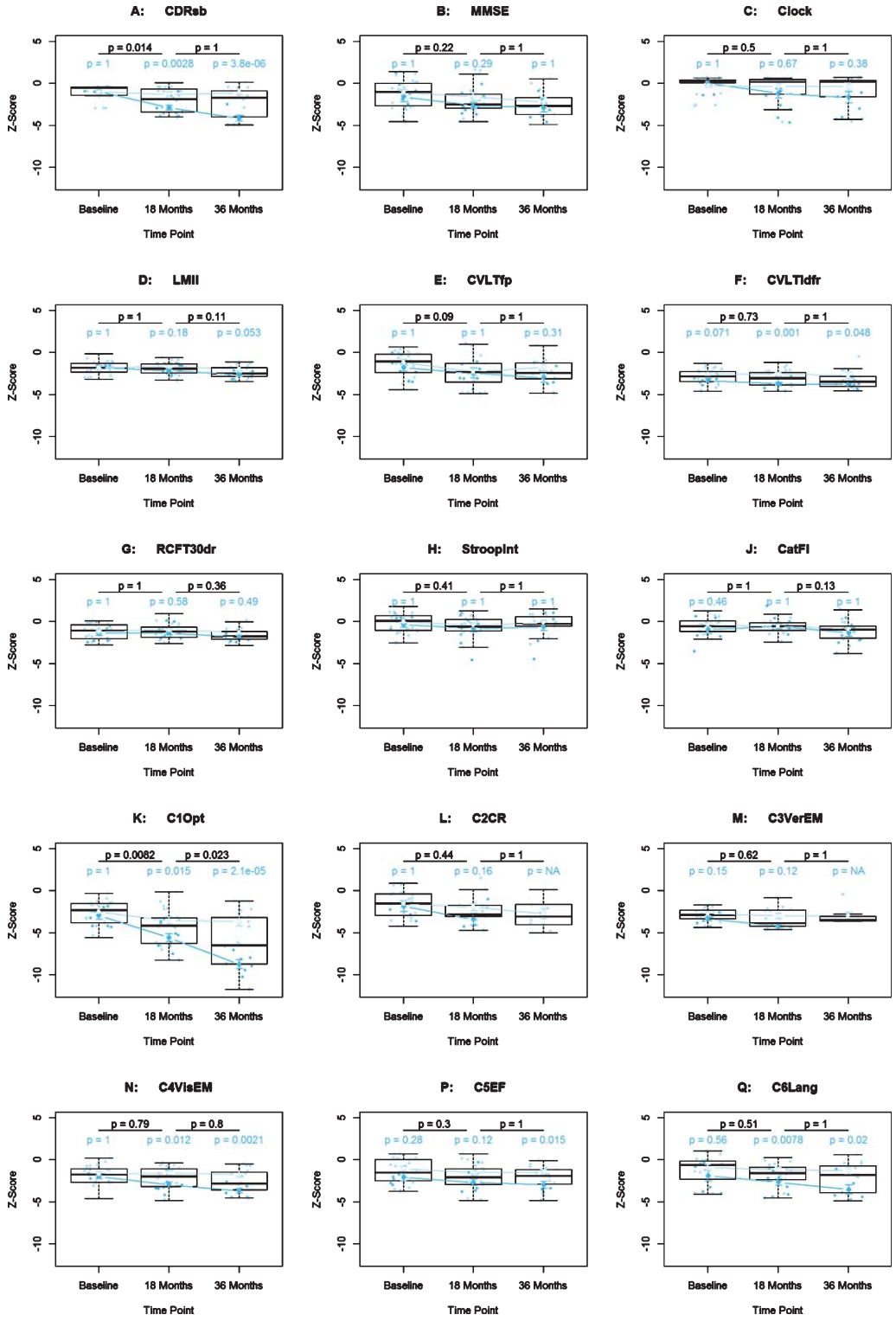

Fig. 2. Longitudinal performance of the standard scores and novel composites. Pale blue data represents individuals who remained stable MCI within the 36 month follow-up period (Non-Progressors). Dark blue data represents individuals who progressed to AD within the 36 month follow-up period (Progressors). *P*-values given in black represent differences between time points for the entire sample (n = 37), where *p*-values given in blue represent differences between Progressors and Non-Progressors (n of 16 and 21, respectively).

These trajectories are split by those who are reclassified as AD within 36 month follow-up ($n = 16$; dark blue) and those who remain classified as MCI at 36 month follow-up ($n = 21$; pale blue). CDRsb (Fig. 2A) appears to capture the most dramatic deterioration with time; however, it can be seen that this is clearly driven by the individuals who are reclassified as AD within 36 months, with the individuals who remain MCI having a very mild decline. A similar profile (although less dramatic) is seen for the Clock score (Fig. 2C). For MMSE (Fig. 2B), the decline (slope) appears to be similar for stable MCIs and those who progress to AD. The measures of episodic memory (Fig. 2D–G) also appear to have similar but shallower profiles to those of MMSE. For $Stroop_{Int}$ (Fig. 2H), there appears to be no difference with time or disease progression. For CatFl (Fig. 2J), there appears to be no decline at 18 months but a mild decline is seen between 18 and 36 months.

Compiling the novel composites

Six composites were compiled using the standard scores evaluated above. One composite, $C1_{Opt}$, was created using the five scores (CDRsb, MMSE, LMII, $CVLT_{FP}$, and Clock) with the highest magnitude F-statistics; the top 5 scores were chosen as there was a natural break in the magnitude of the F-Statistics (a drop of over 20%) between the 5th and 6th ranking scores. A second composite, based on clinical standard scores, $C2_{CR}$, was created by combining CDRsb and MMSE (which also represent the two scores with the highest magnitude of F-statistics). Finally, four domain-specific composites were generated using the three domain-specific scores with the highest magnitude of F-statistics combined with CDRsb: Verbal Episodic Memory ($C3_{VerEM}$), Visual Episodic Memory ($C4_{VisEM}$), Executive Function ($C5_{EF}$), and Language ($C6_{Lang}$). It should be noted that only two domain-specific scores were included in the Language composite as there were only two diverse standard scores for Language in the test battery. The composition of these measures is given in Table 2.

Composite evaluation

Change over time metrics for the novel composites

For the purpose of determining the added efficacy of the clinical/clinical-functional scores of CDRsb and MMSE, six surrogates of the composites identified above were also created. Four of the novel composites (C3-C6) were created without the inclusion of CDRsb (represented by the prefix A). C1 was created without the inclusion of MMSE (represented by the prefix B) and without the inclusion of MMSE or CDRsb (represented by the prefix C).

The magnitude of change from baseline, represented by the F-statistics computed from LME models over

Table 2
Baseline mean and standard deviations (sd) for the norm corrected standard scores and novel composites of interest. Sample size calculations are also given for each measure based on a 25% treatment effect over three years with 80% power. *Raw Scores at baseline are given as Median (IQR) due to the non-normal distribution of the data

Score	Components	Mean raw score at baseline (sd)	Mean corrected Z-score at baseline (sd)	Sample size (95% CI)
CDRsb*	–	0.5 (1.00)	–0.99 (0.79)	162 (161–164)
MMSE	–	26.84 (2.15)	–1.43 (1.89)	418 (413–422)
Clock*	–	10.00 (0.00)	–0.17 (1.04)	664 (653–674)
LMII	–	4.43 (3.23)	–1.81 (0.78)	476 (469–483)
$CVLT_{FP}$	–	8.70 (5.58)	–1.81 (1.89)	423 (417–430)
$CVLT_{LDFR}$	–	3.57 (2.43)	–2.81 (0.9)	1704 (1657–1752)
$RCFT_{30DR}$		8.57 (5.42)	–1.28 (0.93)	1621 (1583–1660)
$Stroop_{Int}$	–	2.45 (0.68)	–0.14 (1.05)	8646 (2819–59447)
CatFl	–	33.32 (9.16)	–0.6 (1.06)	491 (484–498)
$C1_{Opt}$	CDRsb, MMSE, LMII, $CVLT_{FP}$, Clock	–	–2.65 (1.37)	105 (104–106)
B-$C1_{Opt}$	CDRsb, LMII, $CVLT_{FP}$, Clock	–	–2.41 (1.16)	114 (113–115)
C-$C1_{Opt}$	LMII, $CVLT_{FP}$, Clock	–	–1.97 (1.02)	169 (168–171)
$C2_{CR}$	CDRsb, MMSE	–	–2.32 (2.24)	211 (210–213)
$C3_{VerEM}$	CDRsb, LMII, $CVLT_{FP}$, $CVLT_{LDFR}$	–	–3.26 (1.28)	150 (149–152)
A-$C3_{VerEM}$	LMII, $CVLT_{FP}$, $CVLT_{LDFR}$	–	–2.9 (1.19)	310 (305–314)
$C4_{VisEM}$	CDRsb, $RCFT_{3DR}$, $RCFT_{30DR}$, $RCFT_{Hits}$,	–	–1.77 (1.08)	376 (371–380)
$C5_{EF}$	CDRsb, Stroop, FAS, $CatSw_{Tot}$	–	–1.39 (1.4)	348 (344–352)
$C6_{Lang}$	CDRsb, CatFl, BNT	–	–1.55 (2.01)	182 (180–183)

See Supplementary Table 2 for abbreviation definitions.

the 36 month follow-up period, for each of the novel composites and their surrogates are demonstrated by the grey bars in Fig. 1A. It can be seen that the composite with the highest magnitude of F-statistics is the $C1_{Opt}$, and that B-$C1_{Opt}$ (without MMSE) has similar performance. The next highest magnitudes of F-statistics are seen for C- $C1_{Opt}$ (without MMSE or CDRsb) and the $C2_{CR}$ (just MMSE and CDRsb). Three of the domain specific composites ($C3_{VerEM}$, $C5_{EF}$, $C6_{Lang}$) also have some of the higher magnitudes of F-statistics. Eight out of the top 10 measures with the highest magnitude of F-statistics were novel composite measures (CDRsb and MMSE were the standard scores exhibited in the top 10): three of these were muliti-domain/global cognitive-functional composites, three were single domain cognitive-functional composites, one was a multi-domain cognitive composite and one was a single domain cognitive composite.

Cognitive-functional composites (which included CDRsb) always had higher magnitudes of F-statistics than the counterpart composites which did not include CDRsb (prefixes A & C), likewise the composite that did not include MMSE (prefix B) did not have as high a magnitude as its counterpart which did include MMSE.

Similar patterns are also reflected in the 18 and 36 month GLS models (Supplementary Figure 1).

Longitudinal performance of the novel composites

Thirty-six month trajectories for the six novel composites are given in Fig. 2K–Q. For all composites, individuals who progressed to AD within 36 months had lower absolute values. For $C1_{Opt}$, $C4_{VisEM}$, and $C6_{Lang}$ the rates of decline (slopes) appear to be greater in the individuals who progressed to AD within 36 months. For the other composites, there appears to be no difference in the rates of decline. Supplementary Figure 2K–Q details the trajectories of these measures for all disease stages.

Power and sample size calculations

Figure 3 details plots of the statistical power (y-axis) as a function of sample size requirements (x-axis), for a hypothesized 25% treatment effect over three years. These are provided for the nine standard scores and six composites listed in Table 2. Table 2 also summarizes the three year change from baseline and the number (with 95% confidence intervals (CIs)) needed in each arm of a trial to observe a 25% treatment effect over three years with 80% power. It can be seen that $C1_{Opt}$, $C3_{VerEM}$, and CDRsb are the three measures with the greatest advantage for monitoring efficacy over a three year trial. To observe a 25% treatment effect with 80% power 105 participants per arm are required for $C1_{Opt}$, 150 per arm for $C3_{VerEM}$, and 162 per arm for CDRsb.

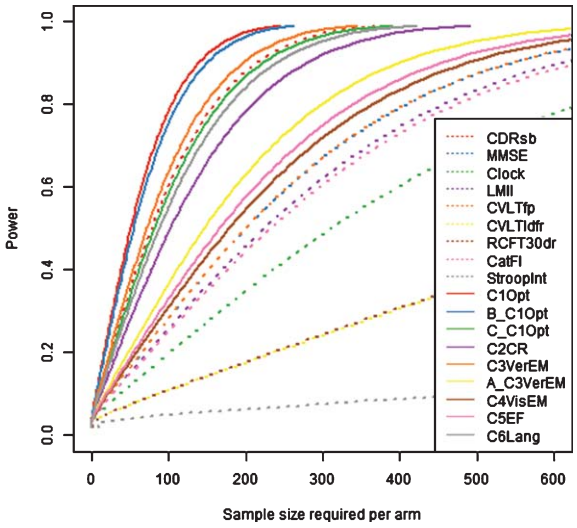

Fig. 3. Power calculation plots: where a solid line represents a novel composite and a dashed line a standard score. From left to right the lines represent $C1_{Opt}$, B_C1_{Opt}, $C3_{VerEM}$, CDRsb, C_C1_{Opt}, $C6_{Lang}$, $C2_{CR}$, A_C3_{VerEM}, $C5_{EF}$, $C4_{VisEM}$, MMSE, $CVLT_{FP}$, LMII, CatFl, Clock, $RCFT_{30DR}$, $CVLT_{LDFR}$, $Stroop_{Int.}$

Rank comparison with overlapping ADNI standard scores

Eight overlapping standard scores were compared, namely: CDRsb, MMSE, LMI, LMII, $CVLT_{LDFR}$, Clock, BNT, and CatFl. For the ADNI dataset, the Rey Auditory Verbal Learning Test (AVLT) [29] long delayed free recall was substituted for $CVLT_{LDFR}$, also CatFl in AIBL utilized Animals and Boys Names whereas in ADNI only Animals was utilized. Further, in ADNI, LMI and LMII were taken from WMS-R [30] opposed to WMS-III for AIBL. The ranking given by the AIBL dataset was CDRsb, MMSE, Clock, LMII, CatFl, $CVLT_{LDFR}$, LMI, and BNT. For the ADNI analyses, the ranking was CDRsb, MMSE, BNT, Clock, LMII, $AVLT_{LDFR}$, CatFl, and LMI. A Spearman's rank correlation coefficient of 0.62 was observed. A table and figure outlining these results are given in Supplementary material (Supplementary Table 1 and Supplementary Figure 3).

DISCUSSION

This study represents a statistical evaluation of standard clinical and neuropsychological measures versus novel composites derived from these parent measures to derive a suitable endpoint measure for prodromal clinical trials. The standard scores and novel composites were evaluated using change from baseline, longitudinal, and power evaluations. The corroboration of this study with findings from other similar studies [5, 31, 32] based on geographically different cohorts adds weight to current thinking in the field. This is further exemplified by the rank comparison validation which demonstrates similar performance in standard tests across the AIBL and ADNI cohorts.

The standard scores consistently demonstrating the largest magnitude of change from baseline (Fig. 1) were CDRsb and MMSE. It appeared that CDRsb captured a short rapid decline exhibited by MCI individuals at the transitional period between MCI and AD; however, it did not readily capture decline in MCI individuals who were not at this boundary (Fig. 2A). Conversely MMSE did not appear to capture different dynamics between MCI individuals who progressed to AD and those who remained MCI over the 36 month follow-up period: the slopes for each group were similar; however, the absolute values differed suggesting greater impairment in the individuals who progressed to AD (Fig. 2B). This may suggest that MMSE provides a more general measure of impairment detectable throughout the MCI spectrum, with CDRsb being more specific to aspects of decline associated with progression to AD. As classification of AD often includes consideration of functional deficit and given that there is a functional aspect to the CDRsb score, this finding is not altogether surprising.

Given that CDRsb and MMSE were the standard scores consistently displaying the largest magnitude of change from baseline, and that their longitudinal dynamics differed from each other, it was hypothesized that they may provide complementary information if compiled into a novel composite, $C2_{CR}$. This novel composite had the second largest magnitude of change from baseline of all the novel composites (Fig. 1). It is only superseded by the combination of these two measures with three additional standard scores (LMII, $CVLT_{FP}$, and Clock) to generate $C1_{Opt}$. Power calculations also demonstrated the added efficacy provided by $C1_{Opt}$, over $C2_{CR}$, with 105 *c.f.* 211 individuals required for each arm of a trial to observe a 25% treatment effect over three years with 80% power.

Four domain-specific cognitive-functional composites were also generated based on domain specific standard scores exhibiting the largest magnitude of change from baseline, as well as CDRsb as a functional measure. These composites were created for the domains of verbal episodic memory, visual episodic memory, executive function, and language. The association of a standard score to the relevant domain was identified from the literature [33]. For each domain, a cognitive-only composite was also generated, without the inclusion of CDRsb (prefix A). The novel cognitive-functional composites always showed increased magnitudes of change in comparison to their cognitive composite counterparts (Fig. 1), suggesting functional deficits are apparent at this stage of the disease course, which are complementary to the cognitive deficits seen. It should also be noted that the multi-domain and clinical novel composites (e.g., mean F-Statistics of 5.39 and 4.57 for $C1_{Opt}$ and $C2_{CR}$, respectively) outperformed the domain specific composites (e.g., mean F-Statistics of 3.91 and 3.07 for $C3_{VerEM}$ and $C5_{EF}$, respectively), and generally the statistically-derived novel composites performed better than the standard scores from which they were derived.

The major limitation associated with this study is the small number of subjects available for analysis; this is somewhat balanced by the very strict criteria placed upon the subjects to obtain a stringent MCI cohort for evaluation and that the results presented here closely match those seen in similar reports [5, 31, 32]. Another potential limitation is that for the AIBL study, the CDR assessors are not blinded to the outcomes of the neuropsychological battery and therefore are making extremely well-informed CDR assessments. It is, therefore, likely that this practice is influencing the strength seen by CDRsb in this study. It should also be noted that the results presented here are based on an observational study and that the true ability of any measure for capturing efficacy of a therapeutic will only be apparent in the setting of a successful clinical trial. It should also be noted that this is not a completers study and that the number of individuals that dropped out or died by 36 month follow-up was six.

In conclusion, an empirically based method was employed to generate novel statistically-derived composite measures from standard test scores, which have improved sensitivity in the assessment of change from baseline. As clinical trials are now focusing on participants earlier in the AD disease course, the novel composites were developed using data from MCI participants with high neocortical Aβ burden in the

AIBL study. Composites comprising measures including function, general cognition and episodic memory appeared to best capture this change. These composites may be useful to assess progression or lack thereof in prodromal AD. However, given the modest sample sizes in this study, these results need to be replicated and validated in a larger independent sample to further test the performance of these derived values as a useful tool. We are encouraged, however, that we observed comparable test performance between AIBL and ADNI participants.

ACKNOWLEDGMENTS

The authors would like to thank all those who took part as participants in the study, as well as clinicians who referred patients with AD and MCI, for their commitment and dedication to helping advance research into the early detection and causation of AD.

Authors' disclosures available online (http://j-alz.com/manuscript-disclosures/14-3015r2).

AIBL: Core funding for the study was provided by the CSIRO Flagship Collaboration Fund and the Science and Industry Endowment Fund (SIEF) in partnership with Florey Institute of Neuroscience and Mental Health, Edith Cowan University (ECU), Alzheimer's Australia (AA), National Ageing Research Institute (NARI), Austin Health, CogState Ltd., Hollywood Private Hospital, Sir Charles Gardner Hospital. The study also receives funding from the National Health and Medical Research Council (NHMRC), the Dementia Collaborative Research Centres program (DCRC), The McCusker Alzheimer's Research Foundation and Operational Infrastructure Support from the Government of Victoria. AIBL has also received funding from industry to support the cohort including Pfizer, GE, Merck and Janssen.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-143015.

REFERENCES

[1] Rosen WG, Mohs RC, Davis KL (1984) A new rating scale for Alzheimer's disease. *Am J Psychiatry* **141**(11), 1356-1364.

[2] Cano SJ, Posner HB, Moline ML, Hurt SW, Swartz J, Hsu T, Hobart JC (2010) The ADAS-cog in Alzheimer's disease clinical trials: Psychometric evaluation of the sum and its parts. *J Neurol Neurosurg Psychiatry* **81**, 1363-1368.

[3] Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, Zajicek J (2013) Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods. *Alzheimers Dement* **9**, S4-S9.

[4] Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, Zajicek J (2013) Putting the Alzheimer's cognitive test to the test II: Rasch Measurement Theory. *Alzheimers Dement* **9**, S10-S20.

[5] Raghavan N, Samtani MN, Farnum M, Yang E, Novak G, Grundman M, Narayan V, DiBernardo A (2013) The ADAS-Cog revisited: Novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. *Alzheimers Dement* **9**, S21-S31.

[6] 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 (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270-279.

[7] Wessels AM, Raghavan N, Yu P (2014) Retrofitting existing tools across the Alzheimer's disease spectrum. *Alzheimers Dement* **10**, 244-245.

[8] Donohue M, Sperling R, Salmon D, Rentz D, Raman R, Thomas R, Weiner M, Aisen P (2014) The preclinical Alzheimer cognitive composite: Measuring amyloid-related decline. *JAMA Neurol* **71**, 961-970.

[9] Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NI, Lenzo N, Martins RN, Maruff P, Masters C, Milner A, Pike K, Rowe C, Savage G, Szoeke C, Taddei K, Villemagne V, Woodward M, Ames D, Grp AR (2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr* **21**, 672-687.

[10] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical-Diagnosis of Alzheimers-Disease - Report of the NINCDS-ADRDA Work Group under the Auspices of Department-of-Health-and-Human-Services Task-Force on Alzheimers-Disease. *Neurology* **34**, 939-944.

[11] Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund LO, Nordberg A, Backman L, Albert M, Almkvist O, Arai H, Basun H, Blennow K, de Leon M, DeCarli C, Erkinjuntti T, Giacobini E, Graff C, Hardy J, Jack C, Jorm A, Ritchie K, van Duijn C, Visser P, Petersen RC (2004) Mild cognitive impairment - beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment. *J Intern Med* **256**, 240-246.

[12] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E (1999) Mild cognitive impairment - Clinical characterization and outcome. *Arch Neurol* **56**, 303-308.

[13] Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI) clinical characterization. *Neurology* **74**, 201-209.

[14] Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J Psychiatric Res* **12**, 189-198.

[15] Delis D, Kramer J, Kaplan E, Ober B (2000) *California Verbal Learning Test-Second Edition*. The Psychological Corporation, San Antonio, TX.

[16] Wechsler D (1945) A standardized memory scale for clinical use. *J Psychol* **19**, 87-95.

[17] Strauss E, Sherman EMS, Spreen O (2006) *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary,* 3rd ed, Oxford University Press, New York.

[18] Wechsler D (1997) *Wechsler Memory Scale—Third Edition*, The Psychological Corporation, San Antonio.

[19] Delis D, Kaplan E, Kramer J. (2001) *Delis-Kaplan Executive Function System (D-KEFS)*. The Psychological Corporation, San Antonio, TX.

[20] Saxton J, Ratcliff G, Munro CA, Coffey EC, Becker JT, Fried L, Kuller L (2000) Normative data on the Boston Naming Test and two equivalent 30-item short forms. *Clin Neuropsychologist* **14**, 526-534.

[21] Meyers J, Meyers K (1995) *Rey Complex Figure Test and Recognition Trial. Professional Manual*, Psychological Assessment Resource, Inc.

[22] Wechsler D (1997) *Wechsler Adult Intelligence Scale, 3rd Edition (WAIS III).* The Psychological Corporation, San Antonio, TX.

[23] Wechsler D (2001) *Wechsler Test of Adult Reading: Examiner's Manual.* The Psychological Corporation, San Antonio, TX.

[24] Burnham S, Raghavan N, Wilson W, Baker D, Ropacki M, Novak G, Ames D, Ellis K, Martins R, Maruff P, Masters C, Romano G, Rowe C, Savage G, Twyman R, Macaulay L, Narayan V (2014) *Comparison of Three Normative Data Correction Approaches: A Cross Sectional Evaluation in the AIBL Study.* Elsevier, Alzheimer's Association International Conference.

[25] Holtzer R, Goldin Y, Zimmerman M, Katz M, Buschke H, Lipton RB (2008) Robust norms for selected neuropsychological tests in older adults. *Arch Clin Neuropsychol* **23**, 531-541.

[26] Rowe CC, Ellis KA, Rimajova M, Bourgeat P, Pike KE, Jones G, Fripp J, Tochon-Danguy H, Morandeau L, O'Keefe G, Price R, Raniga P, Robins P, Acosta O, Lenzo N, Szoeke C, Salvado O, Head R, Martins R, Masters CL, Ames D, Villemagne VL (2010) Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging. *Neurobiol Aging* **31**, 1275-1283.

[27] Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R (2014) Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *Lancet Neurol* **13**, 614-629.

[28] Liu G, Liang K-Y (1997) Sample size calculations for studies with correlated observations. *Biometrics* **53**, 937-947.

[29] Rey A (1958) *L'examen clinique en psychologie*.

[30] Wechsler D (1987) *Wechsler Memory Scale—Revised*, The Psychological Corporation, San Antonio.

[31] Huang Y, Ito K, Billing CB Jr, Anziano RJ (2015) Development of a straightforward and sensitive scale for MCI and early AD clinical trials. *Alzheimers Dement* **11**, 404-414.

[32] Hendrix SB (2012) Measuring clinical progression in MCI and pre-MCI populations: Enrichment and optimizing clinical outcomes over time. *Alzheimers Res Ther* **4**, 24.

[33] Harrington KD, Lim YY, Ellis KA, Copolov C, Darby D, Weinborn M, Ames D, Martins RN, Savage G, Szoeke C, Rowe C, Villemagne VL, Masters CL, Maruff P (2013) The association of A beta amyloid and composite cognitive measures in healthy older adults and MCI. *Int Psychogeriatrics* **25**, 1667-1677.