Original Research

The Journal of Prevention of Alzheimer's Disease - JPAD©
Volume 3, Number 4, 2016

'Alzheimer's Progression Score': Development of a Biomarker Summary Outcome for AD Prevention Trials

J.-M. Leoutsakos[1], A.L. Gross[2], R.N. Jones[3], M.S. Albert[4], J.C.S. Breitner[5]

1. Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA; 2. Departments of Epidemiology and Mental Health, Johns Hopkins Center on Aging and Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; 3. Department of Neurology and Psychiatry & Human Behavior, Warren Alpert Medical School, Brown University, Providence, RI, USA; 4. Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 5. Centre for Studies on Prevention of Alzheimer's Disease (StoP-AD), Douglas Mental Health University Institute Research Centre, McGill University Faculty of Medicine. Montreal, Canada

Corresponding Author: Jeannie-Marie Leoutsakos, 5300 Alpha Commons Drive, Baltimore, MD 21224, USA, jeannie-marie@jhu.edu, Tel: 1-410-550-9884, Fax: 1-410-550-1407

J Prev Alz Dis 2016;3(4):229-235
Published online October 25, 2016, http://dx.doi.org/10.14283/jpad.2016.120

Abstract

BACKGROUND: Alzheimer's disease (AD) prevention research requires methods for measurement of disease progression not yet revealed by symptoms. Preferably, such measurement should encompass multiple disease markers.

OBJECTIVES: Evaluate an item response theory (IRT) model-based latent variable Alzheimer Progression Score (APS) that uses multi-modal disease markers to estimate pre-clinical disease progression.

DESIGN: Estimate APS scores in the BIOCARD observational study, and in the parallel PREVENT-AD Cohort and its sister INTREPAD placebo-controlled prevention trial. Use BIOCARD data to evaluate whether baseline and early APS trajectory predict later progression to MCI/dementia. Similarly, use longitudinal PREVENT-AD data to assess test measurement invariance over time. Further, assess portability of the PREVENT-AD IRT model to baseline INTREPAD data, and explore model changes when CSF markers are added or withdrawn.

SETTING: BIOCARD was established in 1995 and participants were followed up to 20 years in Baltimore, USA. The PREVENT-AD and INTREPAD trial cohorts were established between 2011-2015 in Montreal, Canada, using nearly identical entry criteria to enroll high-risk cognitively normal persons aged 60+ then followed for several years.

PARTICIPANTS: 349 cognitively normal, primarily middle-aged participants in BIOCARD, 125 high-risk participants aged 60+ in PREVENT-AD, and 217 similar subjects in INTREPAD. 106 INTREPAD participants donated up to four serial CSF samples.

MEASUREMENTS: Global cognitive assessment and multiple structural, functional, and diffusion MRI metrics, sensori-neural tests, and CSF concentrations of tau, Aβ42 and their ratio.

RESULTS: Both baseline values and early slope of APS scores in BIOCARD predicted later progression to MCI or AD. Presence of CSF variables strongly improved such prediction. A similarly derived APS in PREVENT-AD showed measurement invariance over time and portability to the parallel INTREPAD sample.

CONCLUSIONS: An IRT-based APS can summarize multi-modal information to provide a longitudinal measure of pre-clinical AD progression, and holds promise as an outcome for AD prevention trials.

Key words: Pre-clinical, multiple outcome modalities, summary outcome measures, disease progression, prevention trials, Alzheimer's disease.

Introduction

Prevention is now a primary aim of therapeutic research in Alzheimer disease (AD) (1). Early AD prevention trials randomized thousands of symptom-free individuals to candidate treatments or placebo, observing them thereafter for differences in dementia incidence (2–4). Results were disappointing. Unfortunately, none of these costly multi-year trials were anticipated by preliminary human data to evaluate their probable success. The objective of this report is to examine a method that can measure disease progression in preclinical AD and thus use preliminary observations to advance trial design.

An early approach to identification of preventive interventions was evaluation of candidate treatments' capacity to prevent progression from Mild Cognitive Impairment to dementia (5). Such trials typically lasted about 5 years and used samples of 500 - 1000 (6, 7). Thus, although more economical than primary prevention trials ("primary" referring only to absence of symptoms), these MCI trials were neither brief nor inexpensive. More importantly, advanced AD pathology is common in MCI patients (8, 9), suggesting that many have an underlying disease state that precludes effective intervention. Accordingly, the field now looks increasingly to interventions in the decades of disease development that precede MCI and AD dementia (10), i.e., in the pre-clinical (or, more precisely, pre-symptomatic) stage of AD. For example, the TOMMORROW trial of pioglitazone has as its primary endpoint time to diagnosis of MCI due to AD, but includes a (continuous) composite cognitive z-score as a secondary outcome (11). The A4 trial of solanezumab (12) and the Alzheimer's Prevention Initiative trial of crenezumab (13) also include among their primary outcomes composite cognitive scores.

While clinically important, cognitive measures alone may be insufficient disease indicators at a stage when cognitive changes may be subtle. Other manifestations of pre-symptomatic AD likely include changes in biochemical, imaging, and sensori-neural measures that

Received August 10, 2016
Accepted for publication September 1, 2016

underlie cognitive decline. For example, cerebrospinal fluid (CSF) Aβ42 shows changes in individuals with dominantly inherited AD many years before onset of cognitive deficits (14). Similar results have been observed among late middle-aged offspring of parents with "sporadic" AD (15), who also appear to have early changes in shape and thickness of cortical structures (16, 17). Likewise, diffusion tensor imaging (DTI) measures in cognitively healthy individuals appear to decline more rapidly among those at elevated risk for AD (18). Comparable results have been observed using odor identification (19), and at least one study has suggested that deficits in central auditory processing may predict imminent AD symptoms (20). To the extent that these various measures can provide a coherent "signal", it seems advantageous to consider them together when evaluating treatment effects. However, we know of no previous attempts to use a multi-modal composite outcome measures for AD prevention trials .

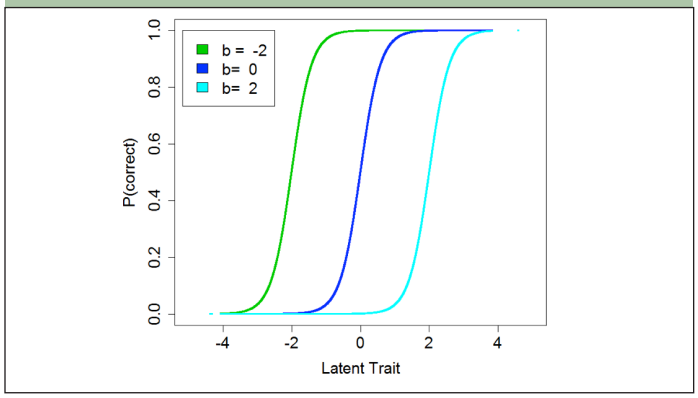
If AD prevention research is to take advantage of the multiplicity of evident pre-symptomatic disease markers, it is essential that methods become available to interpret them in aggregate. To that end, we illustrate the potential of a multi-modal 'Alzheimer Progression Score' (APS) that uses a latent variable modeling approach based on Item Response Theory (IRT) to incorporate data from different modalities to indicate the advance of the pre-symptomatic disease process. We demonstrate this approach in models combining CSF and plasma biochemistries, structural, functional, and diffusion imaging measures, and cognitive and neuro-sensory abilities. We discuss methodologic foundations for this approach and exemplify its potential in several cohorts of individuals at above-average risk of developing AD. Specifically, we present evidence to suggest 1) its construct validity (i.e., that it appears, as desired, to represent degree of advancement in AD pathology), 2) its utility for longitudinal studies (inasmuch as model parameters estimated at baseline can be used to measure disease state over succeeding years of examination), 3) its "portability" to other, similar data sets (useful for clinical trials if one wishes to avoid estimation of model parameters from trial data themselves), and 4) its robustness to missing data, a common occurrence in multi-modal data sets.

Methods

Latent variable modeling as a means of constructing composite measures

Medicine has a long tradition of using composite variables to measure states that are not directly measurable (21). A familiar example is the Framingham Risk Score (22), which provides a global estimate of the 10-year probability of a myocardial infarction (which cannot be measured directly) by assessing several individually measurable risk factors. One approach to construction of such composites involves latent variable methods that model causal relationships between observed variables and hypothetical or unobservable (latent) quantities (23).

Figure 1. Examples of Item Characteristic Curves



Item Response Theory Models

Item response theory (IRT) models are a specific type of latent variable model. In basic IRT models a continuous latent variable (e.g., academic achievement) is presumed to exist and to cause variation in a set of observations (e.g., test question answers). To measure a broad range of the latent variable, the test can be constructed using items that vary in difficulty so that most, including those with lower abilities, can answer the "easier" questions correctly, while only a select few will solve the most difficult. The functional relationship between item responses and the underlying latent variable can be illustrated using item characteristic curves, which are typically sigmoid and identify a region over which an item can usefully discriminate between people with higher and lower levels of the latent trait (Figure 1). A simple IRT model estimates two parameters for each dichotomous (e.g., correct/incorrect, or present/absent) item. These are: a discrimination parameter (a) that describes the slope of the item's response curve and therefore indicates the precision with which it pinpoints an individual's level on the latent variable; and a difficulty/severity parameter (b) that denotes the midpoint of the item's useful range (15). After an IRT model is fit, the resulting parameter estimates can be used to calculate an overall score for each individual.

We use the IRT approach to estimate a latent Alzheimer Progression Score (APS) representing advancement in the evolution of pre-symptomatic AD. The APS "items" are not test questions but measures (e.g., cognitive test scores, CSF protein concentrations, various imaging metrics) that are likely to change as the disease progresses. One advantage of this approach is that the various "items" that contribute to the APS may be sensitive to change in persons who are in earlier or

later stages of disease progression.

Graded response approach

Educational test items typically generate binary scores (right or wrong), but markers of AD pathogenesis are usually continuous, and dichotomization would result in an unacceptable loss of information. Therefore, we use a graded response IRT model (24). Specifically, we transform the continuous data into quintiles (though any quantiles could be used), and the model then estimates discrimination (a) and difficulty/severity (b) parameters for the boundaries between quintiles.

Construct validity of the IRT approach in the BIOCARD Study

We first demonstrated that a summary score, constructed as described, would show construct validity. This meant showing that such a score could measure pre-clinical AD progression and would predict later progression to symptomatic AD. To test this, we needed a sample having a variety of data who were followed for many years. The BIOCARD study met these requirements.

The details of BIOCARD are published elsewhere (25). In brief, the study began in 1995, enrolling 349 cognitively normal middle-aged individuals (mean age 57; SD 10). By design, three quarters of participants had a first degree relative with AD. A neuropsychological battery was administered yearly, and MRI scans, CSF, and blood specimens were obtained at baseline and at biennial follow-up assessments. The study was stopped in 2005, but participants were re-enrolled in 2009, permitting continued follow-up of the cohort. Consensus diagnoses of cognitively normal, MCI, or dementia are now available for up to 20 years of follow-up. BIOCARD was approved by all relevant ethics committees, and all participants provided written informed consent.

We used MPlus version 7.11 (26) to fit a graded response IRT model to BIOCARD's baseline and first two biennial follow-up evaluations using robust maximum likelihood (MLR) estimation. These visits were used because all variables (cognitive, imaging, and CSF) had then been collected. Cognitive indicators included Boston Naming Test (27), logical memory immediate recall and paired associations subscales of the Wechsler Memory Scale-Revised (WMS-R) (28), and the digit symbol substitution subscale of the Wechsler Adult Intelligence Scale-Revised (WAIS-R) (29). MRI measures included right hippocampal and entorhinal cortical volumes, and right entorhinal thickness. CSF measures included Aβ42, total-tau, and phosphorylated tau. These were selected based on evidence of their relationship to progression from normal cognitive status to MCI (16, 25, 30). We transformed items to quintiles (tertiles for Boston Naming Test).

We examined the construct validity of this measurement model by comparing IRT-estimated APS scores for subjects who remained cognitively normal throughout the follow-up with those who progressed to a diagnosis of MCI or dementia. We fit longitudinal mixed effects regression models with random intercepts (31) to compare the two outcome groups on APS scores over time, examining both baseline and rate of change over the three biennial visits. We also estimated the effect of baseline APS score on hazard of MCI/dementia incidence using a cox proportional hazards model (32), and we calculated survival-based area-under-the-ROC-curves (33) using the risksetROC package (34) in R v.3.1.3 (35).

Finally, to assess the information contributed uniquely by the CSF markers, we repeated these procedures after omission of such markers, re-examining the construct validity (predictive capacity) of the resulting, simplified APS scores.

Application of the APS to longitudinal data and a clinical trial

As with any modeling procedure, the IRT-based BIOCARD APS scores relied on the unique combination of data available. Because datasets usually differ in their variables, it is difficult to apply APS scores across studies, unless they incorporate the same data collected using the same methods. The latter similarity of (serial) data collection is the rule in prevention trials, so a single APS model might serve to measure disease progression at multiple timepoints in such trials. To evaluate this, we sought to demonstrate the suitability of an APS as a serial measure in a longitudinal cohort having similar data. Specifically, we looked for evidence that the APS showed measurement invariance over time to provide a valid test of differences in its rate (slope) as an indicator of disease progression in active- vs. placebo-treated groups. This would mean that an APS of, say, "2" meant the same thing at baseline as it did at the end of the study. Demonstration of temporal measurement invariance in a longitudinal cohort would serve our purpose, however, only if the model fit to the longitudinal cohort also fit data from the prevention trial. We refer to this latter attribute as sample portability or invariance. It can be important, typically, if (blinded) trial data are not to be incorporated into the development of the outcome measure used ultimately to compare groups. One may test portability of the method using cohort and trial baseline data (i.e., before treatment assignment). The PREVENT-AD cohort study and its sister INTREPAD trial (NCT 02702817), described in an accompanying paper, presented a fortuitous opportunity for such outcome development.

Temporal Measurement Invariance and Sample Portability in the PREVENT-AD cohort and the INTREPAD trial of naproxen treatment effects in pre-symptomatic AD

PREVENT-AD is an observational cohort of 125 cognitively intact participants. INTREPAD is a two-year, randomized double-blind placebo-controlled trial of naproxen sodium 200 mg b.i.d. for prevention of AD, with a sample size of 217. Except where stated, procedures for the two studies were identical. Participants in both had a parent or multiple siblings affected by probable AD dementia. Participants were aged 60 or older (55 or older if age was within 15 years of their first-affected relative's onset). After careful screening for cognitive disorder, participants were evaluated at baseline and annually using a variety of assessment techniques described elsewhere. INTREPAD participants had an additional visit three months after baseline. Serial CSF samples were not collected in PREVENT-AD, but were collected from 106 (54% of those eligible) INTREPAD enrollees. The trial's principal endpoints are the Repeatable Battery for Assessment of Neuropsychological Status (RBANS) (36) and an APS constructed using other data.

We fit an IRT model and calculated APS scores using baseline data from the PREVENT-AD cohort. Items in this example were UPSIT (University of Pennsylvania Smell Identification Test) (37), cortical thickness in the precuneus, grey matter cerebral blood flow from arterial spin labeling, hippocampal volume, DTI (diffusion tensor imaging) mean diffusivity and fractional anisotropy. These items were chosen based on the current literature linking them to AD progression as well as on observations that they were changing over time.

To assess temporal measurement invariance of the IRT-based APS model for PREVENT-AD, we inquired whether an estimation process that used baseline data only would produce essentially the same parameter estimates as those estimated from all the longitudinal data. Since these models entail a number of parameters, this can be done using a global likelihood ratio test. Specifically, we fit two models to data from all timepoints in the PREVENT-AD Cohort. In one instance, we forced the model parameters (5 for each item; 1 discrimination (a) and 4 difficulty (b) parameters for each quintile boundary) to be the same as the values obtained from a model fit to baseline data only. In the other model we estimated these parameters freely. Depending on sufficiency of information provided (sample size), a likelihood ratio test that was not statistically significant would imply that the unconstrained (freely estimated) model did not fit the data better than the constrained model. Alternately stated, at least within the limits of available data, a null likelihood ratio test would suggest that the IRT model was invariant across time points. We also report the correlation between the individual APS scores resulting from the two models.

The following analogy may be helpful. Consider how a suit of clothing fits a person over time. At baseline a tailor outfits the individual with a suit, after which s/he is fitted with a new suit every year In one of two ways. In the first instance the new suits are made from his/her "baseline" measurements. In the other, the tailor takes new measurements before fitting each new suit. If the person's measurements (analogous to the IRT model) do not change over time, then they show "measurement invariance", and the newly-tailored suits will fit no better than those made using baseline measurements.

Assessment of Portability of IRT model from PREVENT-AD to INTREPAD

We tested the "portability" of the PREVENT-AD IRT model using a similar procedure to that described for temporal invariance. We used only data from the INTREPAD trial baseline, prior to randomization. We fit an IRT model in which we forced parameter estimates to be those obtained from the baseline visit of the PREVENT-AD cohort and compared the fit of that model to one which allowed the parameters to be estimated freely. As before, a null likelihood ratio test suggested portability of the IRT model between the two studies. We also report the correlation between the APS scores from each model.

Assessing the Consequences of Adding CSF to the INTREPAD IRT Model

Given a substantial literature on the relationship between CSF markers and development of AD, we wished to include these markers in the INTREPAD IRT model. However, since these markers were not available in PREVENT-AD, we wished to establish that their addition would not substantially alter the remainder of the parameter estimates (those associated with the other markers), for which we had already demonstrated longitudinal and sample portability. To do so, we compared two models fitted to baseline INTREPAD data, which now included CSF p-tau and Aβ42. In the first model, all non-CSF parameter estimates were constrained to the same values as those from the baseline PREVENT-AD sample, in the second, all parameters were estimated freely. We also report the correlation between the APS scores from each model.

Because a substantial proportion of INTREPAD participants did not have CSF data, we examined how missingness of CSF p-tau and Aβ42 data affected model estimation. Using the subset of trial participants with available CSF, we first fit an IRT model, and then refit it after simulating, randomly, the absence of CSF items for 54% of participants. We then compared the resulting

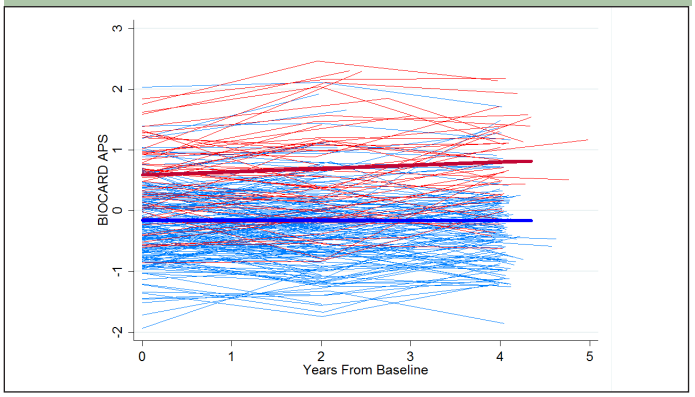
individual APS scores (via correlation) in the complete and simulated missingness data sets.

Results

Findings from BIOCARD

The BIOCARD data used here included 55 who have progressed to MCI, and 22 who have progressed further to dementia. Figure 2 shows individuals' APS over the first, third, and fifth visits, colored by participants' later progression. APS scores are scaled as z-scores. We found that individuals who progressed had significantly higher (worse) scores at baseline (fitted difference: 0.74; SE: 0.09, p<0.001). Individuals who did not progress showed virtually no change in APS score over four years, whereas those who did increased by 0.06 standard deviation units (SE: .02) per year, (p=0.006). In survival analyses, baseline APS was associated with greater hazard of progression over time (HR: 5.76 per SD difference in APS; SE: 1.09; p<0.001). In survival-based ROC analyses the area under the curve (AUC) was 0.80 at ten years, and 0.74 at seventeen years.

Figure 2. APS Scores over Time as a Function of Conversion in BIOCARD



As expected, removal of the CSF data degraded the APS. While those who progressed still had higher scores at baseline (fitted difference: 0.32; SE: 0.11, p=0.003), their difference in annual rate of change from those without progression (0.03; SE: 0.03, vs. 0.01; SE: 0.02) was no longer statistically significant (p=0.447). ROC AUC decreased to 0.61 at ten years and 0.56 at seventeen years.

Findings in PREVENT-AD and INTREPAD

Table 1 shows the similarity of baseline demographics, RBANS total score, and APS items for the PREVENT-AD cohort and INTREPAD trial samples. Table 2 shows parameter estimates from a model using only baseline PREVENT-AD data. For each item, four curves were fitted representing thresholds between each quintiles. Together, the items span a severity range from -3.9 to 5.1 on a z-score scale.

A likelihood ratio test showed that the model assuming temporal measurement invariance did not fit the data significantly worse than a model which estimated all parameters freely (likelihood ratio: 7.593; df:30; p=.999). Likewise, correlation between longitudinal APS scores from IRT models that did and did not assume temporal invariance was 0.97.

Table 1. Baseline Descriptors for PREVENT-AD Cohort and INTREPAD Trial

Variable	Cohort (N=125)	Trial (N=217)
Age (yrs)	63 (5.0)	63 (5.6)
Education (yrs)	16 (3.6)	15 (3.4)
Male	28%	27%
RBANS*	101 (11)	101 (11)
UPSIT†	34.8 (4.4)	35.3 (3.6)
Precuneus Cortical Thickness (mm)	3.2 (0.1)	3.2 (0.1)
Grey Matter Cerebral Blood Flow (mL/110g/min)	53.8 (8.0)	54.8 (8.4)
Hippocampal Volume %‡	0.0002 (0.00002)	0.0002 (0.00003)
DTI§ Mean Diffusivity	3.7 (0.2)	3.7 (0.3)
DTI Fractional Anisotropy	0.4 (0.02)	0.4 (0.02)
CSF|| phosphorylated Tau (pg/mL)	-	47.7 (17.5)
CSF Aβ42 (pg/mL)	-	1063.4 (292.5)

* Repeatable Battery for the Assessment of Neuropsychological Status; † University of Pennsylvania Smell Identification Test; ‡ Hippocampal volume divided by Intracranial volume x 100%; § Diffusion Tensor Imaging; || Cerebrospinal fluid; note that CSF was only collected from 106 trial participants

We also found that the IRT model was portable (sample invariant) between the baseline PREVENT AD cohort and baseline INTREPAD samples (likelihood ratio: 40.613; df: 30; p=.094). The correlation between trial baseline APS scores from the two models was 0.97.

Finally, we determined that addition of CSF P-tau and Aβ42 concentrations to the INTREPAD APS did not significantly alter the model parameters for the original set of items (likelihood ratio: 38.582, df: 30, p=.136). Correlation between APS scores from the two models was 0.94. When assessing the effect of missingness of CSF variables, we found that APS scores calculated from the dataset with simulated missing data had a high (r=0.98) correlation with APS scores calculated from the intact dataset.

Discussion

We describe estimation of an IRT-based 'Alzheimer Progression Score' (APS), a summary measure of pre-symptomatic Alzheimer's disease severity, drawing on a range of available markers of different types (cognitive testing, imaging, CSF chemistries, etc). A principal use of this measure may be as an outcome for clinical trials of interventions intended to slow progression, and thus delay (i.e., prevent) symptom onset. To demonstrate the

Table 2. Parameter Estimates from Baseline PREVENT-AD APS

Variable	Discrimination*	Threshold 1†	Threshold 2	Threshold 3	Threshold 4
UPSIT‡	0.87	-0.51	0.21	0.87	2.21
Precuneus Cortical Thickness	0.31	-3.87	-0.78	1.67	5.08
Grey Matter CBF§	0.52	-2.55	-0.85	1.02	3.34
Hippocampal Volume Fraction	0.74	-1.70	-0.25	0.99	2.65
DTI‖ Mean Diffusivity	1.32	-1.22	-0.13	0.68	1.51
DTI Fractional Anisotropy	1.84	-0.83	0.00	0.66	1.46

* Discrimination represents the precision with which the item measures an individual's progression; † Difficulty/severity parameter at the first threshold (border between quintiles 1 and 2); ‡ University of Pennsylvania Smell Identification Test; § Cerebral blood flow by arterial spin labelling; ‖ Diffusion Tensor Imaging

potential of this approach, we applied it to data from the BIOCARD study. The construct validity of the resulting measure was supported by the fact that individuals who later developed MCI or dementia due to AD had higher APS means at baseline and steeper slopes over time.

For the APS to be a useful outcome measure in clinical trials, we needed also to demonstrate measurement invariance over time, as well as portability (sample invariance) between samples. We were able to demonstrate both features using baseline data from the INTREPAD clinical trial and longitudinal data from the sister PREVENT-AD cohort. Because work in BIOCARD showed that CSF variables added substantial predictive ability to the APS, we tested and demonstrated invariance of the remaining model parameters when CSF biomarker data were added from certain individuals only. We also demonstrated the robustness of the method to missing data from part of the sample -- the latter being particularly important because CSF is commonly available from only a portion of a longitudinal study.

This work addresses several potential concerns about the utility of summary outcome measures such as the APS. First among these is construct validity, i.e., are we truly measuring pre-symptomatic AD severity? At present, our evidence derives from work in BIOCARD, where we showed that an APS using its data (which were considerably less rich than those now available in PREVENT-AD or elsewhere) was a strong predictor of later emergence of MCI or dementia due to AD. An interesting collateral finding from BIOCARD was the range of baseline APS measures from individuals who were all cognitively 'normal', and that such baseline variability was itself strongly predictive of subsequent symptoms. It is unknown whether the higher APS scores in some BIOCARD participants represented deleterious change from an earlier time, but this seems likely. If so, it suggests that at least some differences within the 'normal' range may in fact presage more extreme later change (38). This idea is reinforced in part by the fact that all constituent items of the APS (at least in BIOCARD) have been independently associated with the emergence of AD pathology.

This work has some notable strengths. Chief among them are the parallel design of the PREVENT-AD cohort and INTREPAD trial. This design allowed for the the validation of our measure while still incorporating state-of-the-art items. The use of IRT models to construct composite scores is well-suited to our purpose of measuring pre-clinical AD progression, a continuous latent variable. While relatively new to medicine, these models have a long history in other fields including psychology and education. As such, their properties are relatively well-developed and well-characterized.

We also note a number of limitations. Because of limited duration of follow-up in the other samples, we were able to demonstrate construct validity of an APS only in BIOCARD (though we note that the model in PREVENT-AD drew upon items linked to AD that appeared to change over time. The graded response IRT model that we used entails the quantilization of items that are continuous, an approach that can result in loss of information. While it is possible to re-parameterize a factor analytic model with continuous items as an IRT model (39), it would impose the (unsustainable) assumption that the relationship between AD progression and each item was linear. It is unfortunate that CSF was not collected in the PREVENT-AD cohort. Such inclusion would have permitted better assessment of its potential as part of the INTREPAD APS. Instead, we rely on evidence from the literature, performance in BIOCARD, and the demonstration that its inclusion did not alter the model parameters of the other constituent items in the IRT model fit to baseline INTREPAD data.

Whatever its merits, we recognize that the APS is only an interim solution to the problem of multi-modal data used in estimation of pre-symptomatic disease progression and its treatment. Future directions for this work include finalizing the composition of the INTREPAD APS prior to unblinding of the data. Other methods are also in development (40), and still others are on the horizon (41, 42). Our present purpose is to enable an appropriate analysis of the substantial PREVENT-AD and INTREPAD data sets. We intend eventually to enable public access to these data so that other methods may be tested by others to the same ends.

Acknowledgements: We would like to thank the participants and members of the BIOCARD study. The BIOCARD Study consists of 7 Cores with the following members: (1) the Administrative Core (Marilyn Albert, Barbara Rodzon), (2) the Clinical Core (Ola Selnes, Marilyn Albert, Rebecca Gottesman, Ned Sacktor, Guy McKhann, Scott Turner, Leonie Farrington, Maura Grega, Daniel D'Agostino, , Gay Rudow, Scott Rudow), (3) the Imaging Core (Michael Miller, Susumu Mori, Tilak Ratnanather, Timothy Brown, Hayan Chi, Anthony Kolasny, Kenichi

Oishi, Laurent Younes), (4) the Biospecimen Core (Richard O'Brien, Abhay Moghekar, Richard Meehan), (5) the Informatics Core (Roberta Scherer, David Shade, Ann Ervin, Jennifer Jones, Matt Toepfner, Lauren Parlett, April Patterson, Lisa Lassiter), the (6) Biostatistics Core (Mei-Cheng Wang, Yi Lu, Qing Cai), and (7) the Neuropathology Core (Juan Troncoso, Barbara Crain, Olga Pletnikova, Gay Rudow, Karen Fisher). We would like to acknowledge the contributions to BIOCARD of the Geriatric Psychiatry Branch (GPB) of the intramural program of the NIMH who initiated the study (PI: Dr. Trey Sunderland). The PREVENT AD cohort and INTREPAD trial are key activities of the Centre for Studies on Prevention of Alzheimer's Disease (StoP-AD), Douglas Mental Health University Institute Research Centre. The Centre is co-directed by Dr. Breitner and Drs. Judes Poirier and Pierre E. Etienne. Dr. Pedro Rosa-Neto has performed over 350 lumbar punctures for INTREPAD. Mme. Anne Labonté provided superb technical assistance, and Mmes. Jennifer Tremblay-Mercier and Joanne Frenette provided invaluable coordination of clinical activities. Mme. Melissa Savard worked closely with Dr. Leoutsakos to provide various PREVENT-AD and INTREPAD data for analysis. Mmes. Marie-Elyse Lafaille-Magnan and Cecile Madjar coordinated the olfactory identification and imaging data for inclusion in the A.P.S. models. Other investigators and staff are listed as part of the accompanying manuscript. Most of all, we are grateful for the tireless commitment of our participants who have returned for repeated annual assessments and, in some instances, up to five serial lumbar punctures.

Funding: The BIOCARD study is funded by the National Institutes of Health grant U19-AG033655 (PI Marilyn Albert). PREVENT-AD and INTREPAD are funded by generous support from McGill University, by an unrestricted gift from Pfizer Canada, and by infrastructure support from the Canada Fund for Innovation. Dr. Breitner's effort is supported by a Canada Research Chair award from the government of Canada. Support for our genetic and laboratory work has been provided by the Fonds de Récherche du Québec - Santé (FRQ-S) and by the Levesque Foundation. Imaging work has received support from the FRQ-S. Additional support has been provided by the Douglas Mental Health University Institute Foundation.

Conflict of interest: None.

Ethical standards: The BIOCARD, PREVENT-AD, and INTREPAD studies were conducted under protocols and informed consent procedures approved by all relevant IRBs. Written informed consent was obtained from each participant for all phases of these protocols.

References

1. Reiman EM, Langbaum JBS, Fleisher AS, Caselli RJ, Chen K, Ayutyanont N, et al. Alzheimer's Prevention Initiative: a plan to accelerate the evaluation of presymptomatic treatments. J Alzheimers Dis 2011;26 Suppl 3:321–9.
2. Sano M, Jacobs D, Andrews H, Bell K, Graff-Radford N, Lucas J, et al. A multi-center, randomized, double blind placebo-controlled trial of estrogens to prevent Alzheimer's disease and loss of memory in women: design and baseline characteristics. Clin Trials 2008;5:523–33.
3. DeKosky ST, Williamson JD, Fitzpatrick AL, Kronmal RA, Ives DG, Saxton JA, et al. Ginkgo biloba for prevention of dementia: a randomized controlled trial. JAMA 2008;300:2253–62.
4. ADAPT Research Group, Lyketsos CG, Breitner JCS, Green RC, Martin BK, Meinert C, et al. Naproxen and celecoxib do not prevent AD in early results from a randomized controlled trial. Neurology 2007;68:1800–8.
5. Jelic V, Kivipelto M, Winblad B. Clinical trials in mild cognitive impairment: lessons for the future. J Neurol Neurosurg Psychiatry 2006;77:429–38.
6. Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. N Engl J Med 2005;352:2379–88.
7. Thal LJ, Ferris SH, Kirby L, Block GA, Lines CR, Yuen E, et al. A randomized, double-blind, study of rofecoxib in patients with mild cognitive impairment. Neuropsychopharmacology 2005;30:1204–15.
8. Haroutunian V, Hoffman LB, Beeri MS. Is there a neuropathology difference between mild cognitive impairment and dementia? Dialogues Clin Neurosci 2009;11:171–9.
9. Schneider JA, Arvanitakis Z, Leurgans SE, Bennett DA. The neuropathology of probable Alzheimer disease and mild cognitive impairment. Ann Neurol 2009;66:200–8.
10. Dubois B, Hampel H, Feldman HH, Scheltens P, Aisen P, Andrieu S, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. Alzheimers Dement 2016/3;12:292–323.
11. Biomarker Qualification for Risk of Mild Cognitive Impairment (MCI) Due to Alzheimer's Disease (AD) and Safety and Efficacy Evaluation of Pioglitazone in Delaying Its Onset - Full Text View - ClinicalTrials.gov n.d. https://clinicaltrials.gov/ct2/show/NCT01931566?term=tommorrow&rank=1 (accessed May 24, 2016).
12. Sperling RA, Rentz DM, Johnson KA, Karlawish J, Donohue M, Salmon DP, et al. The A4 study: stopping AD before symptoms begin? Sci Transl Med 2014;6:228fs13.
13. A Study of Crenezumab Versus Placebo in Preclinical PSEN1 E280A Mutation Carriers to Evaluate Efficacy and Safety in the Treatment of Autosomal-Dominant Alzheimer Disease (AD), Including a Placebo-Treated Noncarrier Cohort - Tabular View - ClinicalTrials.gov n.d. https://clinicaltrials.gov/ct2/show/record/NCT01998841?term=crenezumab&rank=3 (accessed May 24, 2016).
14. Bateman RJ, Xiong C, Benzinger TLS, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med 2012;367:795–804.
15. Sutphen CL, Jasielec MS, Shah AR, Macy EM, Xiong C, Vlassenko AG, et al. Longitudinal Cerebrospinal Fluid Biomarker Changes in Preclinical Alzheimer Disease During Middle Age. JAMA Neurol 2015;72:1029–42.
16. Younes L, Albert M, Miller MI, BIOCARD Research Team. Inferring changepoint times of medial temporal lobe morphometric change in preclinical Alzheimer's disease. Neuroimage Clin 2014;5:178–87.
17. Donix M, Burggren AC, Suthana NA, Siddarth P, Ekstrom AD, Krupa AK, et al. Longitudinal changes in medial temporal cortical thickness in normal subjects with the APOE-4 polymorphism. Neuroimage 2010;53:37–43.
18. Rieckmann A, Van Dijk KRA, Sperling RA, Johnson KA, Buckner RL, Hedden T. Accelerated decline in white matter integrity in clinically normal individuals at risk for Alzheimer's disease. Neurobiol Aging 2016;42:177–88.
19. Growdon ME, Schultz AP, Dagley AS, Amariglio RE, Hedden T, Rentz DM, et al. Odor identification and Alzheimer disease biomarkers in clinically normal elderly. Neurology 2015;84:2153–60.
20. Gates GA, Anderson ML, McCurry SM, Feeney MP, Larson EB. Central auditory dysfunction as a harbinger of Alzheimer dementia. Arch Otolaryngol Head Neck Surg 2011;137:390–5.
21. Feinstein AR. Multi-item Instruments vs Virginia Apgar's Principles of Clinimetrics. Arch Intern Med 1999;159:125–8.
22. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–47.
23. Gross AL, Mungas DM, Leoutsakos JS, Albert M, & Jones RN. Alzheimer's disease severity, objectively determined and measured. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, in press.
24. Samejima F. Graded response model of the latent trait theory and tailored testing. Proceedings of the first conference on computerized adaptive testing, ERIC; 1976, p. 5–17.
25. Moghekar A, Li S, Lu Y, Li M, Wang M-C, Albert M, et al. CSF biomarker changes precede symptom onset of mild cognitive impairment. Neurology 2013;81:1753–8.
26. Muthén LK, Muthén BO. Mplus User's Guide: Statistical Analysis with Latent Variables : User'ss Guide. Muthén & Muthén; 2010.
27. Katsumata Y, Mathews M, Abner EL, Jicha GA, Caban-Holt A, Smith CD, et al. Assessing the discriminant ability, reliability, and comparability of multiple short forms of the Boston Naming Test in an Alzheimer's disease center cohort. Dement Geriatr Cogn Disord 2015;39:215–27.
28. Wechsler D. Wechsler Memory Scale-Revised Manual (The Psychological Corporation, San Antonio, TX) 1987.
29. Wechsler D. Manual for the adult intelligence scale-revised. New York: Psychological Corporation 1981.
30. Soldan A, Pettigrew C, Li S, Wang M-C, Moghekar A, Selnes OA, et al. Relationship of cognitive reserve and cerebrospinal fluid biomarkers to the emergence of clinical symptoms in preclinical Alzheimer's disease. Neurobiol Aging 2013;34:2827–34.
31. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–74.
32. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. Wiley; 2011.
33. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337–44.
34. CRAN - Package risksetROC n.d. https://cran.r-project.org/web/packages/risksetROC/index.html (accessed May 24, 2016).
35. R Core Team. R: A Language and Environment for Statistical Computing 2014.
36. Randolph C, Tierney MC, Mohr E, Chase TN. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity. J Clin Exp Neuropsychol 1998;20:310–9.
37. Doty RL, Shaman P, Kimmelman CP, Dann MS. University of Pennsylvania Smell Identification Test: a rapid quantitative olfactory function test for the clinic. Laryngoscope 1984;94:176–8.
38. Jagust W. Vulnerable neural systems and the borderland of brain aging and neurodegeneration. Neuron 2013;77:219–34.
39. Kamata A, Bauer DJ. A Note on the Relation Between Factor Analytic and Item Response Theory Models. Struct Equ Modeling 2008;15:136–53.
40. Iturria-Medina Y, Sotero RC, Toussaint PJ, Mateos-Pérez JM, Evans AC, The Alzheimer's Disease Neuroimaging Initiative. Early role of vascular dysregulation on late-onset Alzheimer/'s disease based on multifactorial data-driven analysis. Nat Commun 2016;7. doi:10.1038/ncomms11934.
41. Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, et al. A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort. Neuroimage 2012;63:1478–86.
42. Xiong C, van Belle G, Chen K, Tian L, Luo J, Gao F, et al. Combining Multiple Markers to Improve the Longitudinal Rate of Progression: Application to Clinical Trials on the Early Stage of Alzheimer's Disease. Stat Biopharm Res 2013;5:54–66.