Neuro
epidemiology

Neuroepidemiology 2014;43:194–205
DOI: 10.1159/000367970

Received: July 1, 2014
Accepted: August 23, 2014
Published online: November 13, 2014

# Calibrating Longitudinal Cognition in Alzheimer's Disease Across Diverse Test Batteries and Datasets

Alden L. Gross[a] Richard Sherva[b] Shubhabrata Mukherjee[c] Stephen Newhouse[d] John S.K. Kauwe[e] Leanne M. Munsie[f] Leo B. Waterston[g] David A. Bennett[h] Richard N. Jones[i] Robert C. Green[g] Paul K. Crane[c] for the Alzheimer's Disease Neuroimaging Initiative, GENAROAD Consortium, and AD Genetics Consortium

[a]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Md., [b]Boston University School of Medicine, Boston, Mass., [c]Department of Medicine, University of Washington, Seattle, Wash., USA; [d]King's College London, Institute of Psychiatry, London, NIHR Biomedical Research Centre for Mental Health at South London and Maudsley NHS Foundation, London, UK; [e]Departments of Biology and Neuroscience, Brigham Young University, Provo, Utah, [f]Tailored Therapeutics, Eli Lilly and Company, Indianapolis, Ind., [g]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Mass., [h]Rush Alzheimer's Disease Center, Department of Neurological Sciences, Rush University Medical Center, Chicago, Ill., [i]Departments of Psychiatry and Human Behavior and Neurology, Warren Alpert Medical School, Brown University, Providence, R.I., USA

**Key Words**
Calibration · Neuropsychological performance · Alzheimer's disease

**Abstract**
***Background:*** We sought to identify optimal approaches by calibrating longitudinal cognitive performance across studies with different neuropsychological batteries. ***Methods:*** We examined four approaches to calibrate cognitive performance in nine longitudinal studies of Alzheimer's disease (AD) (n = 10,875): (1) common test, (2) standardize and average available tests, (3) confirmatory factor analysis (CFA) with continuous indicators, and (4) CFA with categorical indicators. To compare precision, we determined the minimum sample sizes needed to detect 25% cognitive decline with 80% power. To compare criterion validity, we correlated cognitive change from each approach with 6-year changes in average cortical thickness and hippocampal volume using available MRI data from the AD Neuroimaging Initiative. ***Results:*** CFA with categorical indicators required the smallest sample size to detect 25% cognitive decline with 80% power (n = 232) compared to common test (n = 277), standardize-and-average (n = 291), and CFA with continuous indicators

R.C. Green and P.K. Crane contributed equally as senior authors.
Please see the end of the manuscript for a list of members of the GENAROAD Consortium.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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.



Alden L. Gross
Johns Hopkins Center on Aging and Health and Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health
2024 E. Monument St., Baltimore, MD 21205 (USA)
E-Mail agross14@jhu.edu

(n = 315) approaches. Associations with changes in biomarkers changes were the strongest for CFA with categorical indicators. ***Conclusions:*** CFA with categorical indicators demonstrated greater power to detect change and superior criterion validity compared to other approaches. It has wide applicability to directly compare cognitive performance across studies, making it a good way to obtain operational phenotypes for genetic analyses of cognitive decline among people with AD.


## Introduction

Genome-wide genetic studies and other large-scale studies require high-quality phenotypes and large samples to be adequately powered. Necessary sample sizes are generally beyond what is available in most existing studies [1]. Developing consortia comprising multiple epidemiologic studies to address genetic questions around particular phenotypes is common because of enhanced power to detect associations [2–4]. In particular, the Genetic Architecture of Rate of Alzheimer's Decline (GENAROAD) consortium was formed to examine the genetic basis of cognitive decline among persons with Alzheimer's disease (AD). Combining samples introduces phenotypic heterogeneity, which is a major challenge to successful genetic studies [5, 6]. In dementia research, growing recognition of the need to leverage existing data sources culminated in the International Database for Longitudinal Studies on Aging and Dementia (IDAD) [7, 8]. In the setting of measuring cognitive decline across different longitudinal studies, which employed a range of different cognitive tests, a common cognitive metric is an important phenotype [9, 10]. There is great interest in developing efficient instruments to measure cognitive performance and change for clinical trials [e.g., 11]. In this study, we empirically evaluated four alternative approaches for calibrating summary scores for cognitive performance across studies of AD that used different batteries of cognitive tests.

Although neuropsychological batteries are ubiquitous in clinical and epidemiologic studies of cognitive aging, there is no single, widely used method of assessing general cognitive performance [12]. This diversity complicates the synthesis of findings across studies. Psychometrically sound common cognitive measures are a centerpiece of the NIH Toolbox initiative [13]. While such standardization efforts may produce more closely aligned datasets in the future, they do not address the need to evaluate existing data collected using different cognitive test batteries.

The most obvious approach would be to use one or more tests that are in common across studies. For example, in 2005 Alzheimer's Disease Research Centers agreed to administer a common battery of tests on all participants through the Uniform Data Set (UDS) initiative [14]. A second approach involves standardizing each test score in a battery by centering on the sample-specific mean and dividing by the sample-specific standard deviation, and then summing or averaging them together. A third approach is to use the confirmatory factor analysis (CFA) to derive summary factors while treating each of the individual test scores as a continuous, linear indicator. The fourth approach to calibrating cognitive performance is to estimate a CFA for cognitive performance, but treat each test as a categorical indicator.

In this study, we used these four approaches to derive summary scores for longitudinal cognitive performance across nine datasets of persons with AD. For each approach, we calculated and compared the minimum sample size required to detect a 25% annual decline in cognitive performance with 80% power. To evaluate criterion validity, we then compared the strength of the association between cognitive change characterized by each approach and changes over up to six years in biologically based markers available from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is one of the studies we analyzed. We chose to examine changes in average cortical thickness and hippocampal volume because cognitive performance is strongly associated with changes in these biomarkers [15–18]. We hypothesized that all approaches would demonstrate significant associations with these biomarkers, but that scores from a categorical indicator CFA would provide the most precise estimates of decline due to better fit to all available cognitive data, and thus greater power to detect associations.

## Methods

*Participants*

Participants come from nine cohort studies and clinical trials of older adults, each of which included participants with AD. Descriptions of each study are in table 1. The studies include ADNI (ADNI1, ADNI-GO, and ADNI2) [19], the Rush Memory and Aging Project (MAP) [20], the Religious Orders Study (ROS) [21], the Cache County Study on Memory and Aging [22], the Myriad Tarenflurbil phase III clinical trial [23], the Lilly Semagacestat phase III trial [24], the AddNeuroMed multicenter European study [25], Adult Changes in Thought (ACT) [26], and the Uniform Data Set from the National Alzheimer's Disease Coordinat-

**Table 1.** Demographic characteristics of the overall and study-specific samples (n = 10,875)

| | Description of study sample | Sample size | Age, years (mean ± SD) | Sex, female n (%) | Race, white n (%) | Education | | | Number of study visits, median (IQR) | Follow-up time (years), median (IQR) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | high school or less | college | graduate | | |
| Full sample | | 10,875 | 77.3±7.6 | 6,017 (55.0) | 9,463 (86.4) | 4,137 (39.4) | 4,282 (40.8) | 2,070 (19.7) | 3.0 (2.0–6.0) | 1.6 (1.0–3.6) |
| Range in sample | | | 60.0–110.0 | | | | | | 1.0–17.0 | 0.0–19.0 |
| AD Neuroimaging Initiative (ADNI) | Clinical and cognitive biomarkers in AD; participants representative of clinical trial samples | 431 | 75.4±6.7 | 175 (40.6) | 401 (93.0) | 98 (22.7) | 197 (45.7) | 136 (31.6) | 5.0 (4.0–6.0) | 2.0 (1.5–4.0) |
| Rush Memory and Aging Project (MAP) | Genetic and environmental risk factors for dementia in a diverse sample | 412 | 82.8±5.9 | 279 (67.7) | 394 (95.6) | 146 (35.4) | 178 (43.2) | 88 (21.4) | 2.0 (1.0–3.0) | 1.0 (0.0–3.8) |
| Religious Orders Study (ROS) | Study of dementia in members of Catholic religious orders | 492 | 78.4±7.1 | 332 (67.8) | 451 (92.2) | 37 (7.5) | 138 (28.1) | 316 (64.4) | 2.0 (1.0–4.0) | 2.2 (0.0–6.0) |
| National Alzheimer's Coordinating Center (NACC) | Coordinating center for Alzheimer's Disease Centers | 5,475 | 77.3±7.9 | 2,924 (53.4) | 4,452 (81.3) | 2,045 (37.6) | 2,044 (37.6) | 1,354 (24.9) | 3.0 (2.0–4.0) | 2.2 (1.1–3.7) |
| Cache County Study on Memory Health and Aging (Cache) | Study of genetic and environmental risk factors for AD | 311 | 84.6±6.5 | 207 (66.6) | 311 (100.0) | 161 (51.9) | 108 (34.8) | 41 (13.2) | 3.0 (2.0–6.0) | 1.9 (1.3–4.9) |
| Myriad Tarenflurbil phase III clinical trial | Phase III clinical trial participants with mild AD from 133 centers | 2,370 | 75.8±7.1 | 1,201 (50.7) | 2,114 (89.2) | 1,103 (46.5) | 1,267 (53.5) | 0 (0.0) | 9.0 (6.0–9.0) | 1.5 (1.0–1.6) |
| Lilly Semagacestat phase III clinical trial | Phase III clinical trial of patients with AD | 350 | 74.1±7.8 | 184 (52.6) | 350 (100.0) | 170 (48.6) | 130 (37.1) | 50 (14.3) | 4.0 (3.0–4.0) | 1.3 (0.9–1.5) |
| AddNeuroMed | Study of biomarkers for AD in the UK | 284 | 77.8±6.5 | 178 (62.7) | 277 (97.5) | 226 (80.1) | 42 (14.9) | 14 (5.0) | 5.0 (3.0–5.0) | 1.0 (1.0–1.0) |
| Adult Changes in Thought (ACT) | Prospective longitudinal cohort study of older adults from a Seattle-area Health Maintenance Organization | 750 | 77.2±6.4 | 487 (64.9) | 688 (91.7) | 321 (42.8) | 308 (41.1) | 121 (16.1) | 6.0 (4.0–7.0) | 8.0 (6.0–12.0) |

SD = Standard deviation; MCI = mild cognitive impairment; CIND = cognitive impairment without dementia.

ing Center (NACC), which includes data from 34 past and present AD Centers [27]. Each cohort recruited participants with and without AD, but for the present study, we restricted the samples to participants and study visits at which AD was diagnosed.

*Four Approaches for Calibrating Cognitive Performance Across Studies*

The first approach, which is to take tests in common across studies, is a 'least common denominator' approach that relies on studies having one or more common tests, and discards potentially informative data about cognitive performance provided by other tests.

Second is the standardize and average approach (fig. 1a) in which all tests are standardized to the same scale and averaged together. This approach is intuitive and succeeds in placing cognitive performance on a common scale in a single sample, but it does not allow differential weighting of tests within a study and fails to ensure equal calibrations of the same test across different studies.

The third approach is using confirmatory factor analysis (CFA) with continuous indicators (fig. 1b) in which tests may be weighted differently, and test means and standard deviations are not fixed to be equal to each other within a study. This may improve measurement precision and also the statistical power, and can be used to calibrate tests into common factors across studies with different batteries. The effect of any particular test on the overall score should be the same across studies with different numbers of tests. This approach is more flexible than the standardize and average approach, but makes a strong assumption of linearity between each cognitive test and the underlying general factor measured by all the tests together. Violations of this assumption are to be expected, as there is no reason to assume that the distance between each score increment on any test will be the same across the range of the underlying trait. For example, decline in the Mini-Mental State Examination (MMSE) from 30 to 28 suggests a more severe decline than decline from 10 to 8 (28–31). Nonlinear relationships

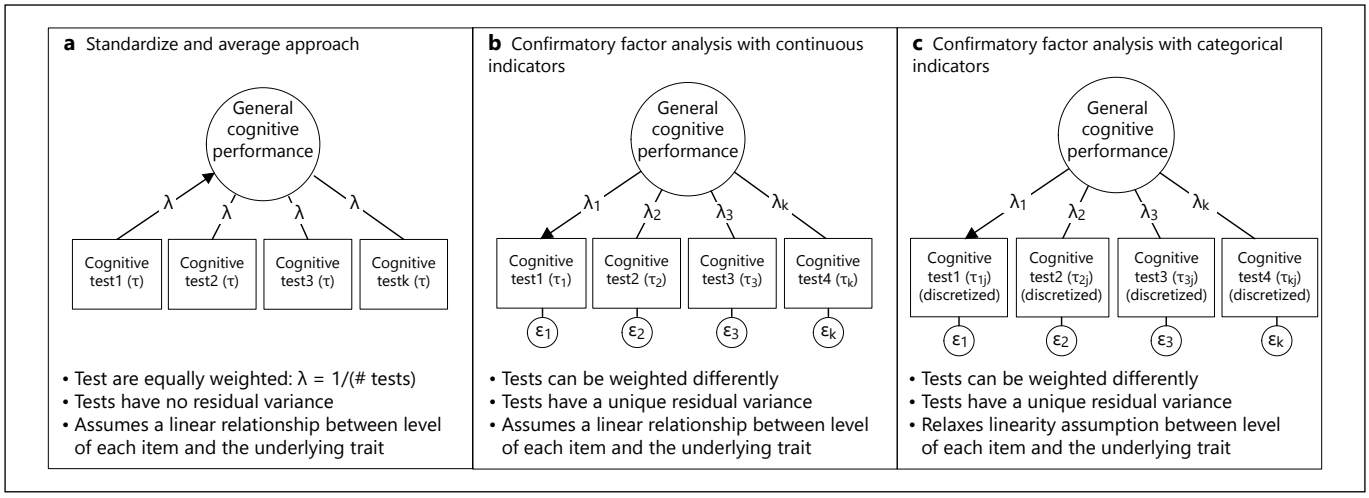


**Fig. 1.** Schematic representation of measurement models for different approaches to deriving summary cognitive scores. This figure compares and contrasts three measurement models for the approaches tested in this study that involve combining multiple tests. Observed cognitive test scores are in squares. Unmeasured constructs, which include the general component and residual variance terms ε, are in circles. Unmeasured general cognitive performance constructs are related to observed tests by factor loadings or weights, λ. Tests have thresholds, τ, that characterize the location along the latent trait where the test provides information. See Methods for details of each approach.

between test scores and the underlying ability, measured by all tests together, pose a challenge to the continuous CFA approach.

The fourth approach entails CFA with categorical indicators (fig. 1c). Discretizing test scores does not require the strong linearity assumption made by the CFA with the continuous indicators approach. Discretization may be viewed as a limitation because it represents a loss of information. However, as performance on many cognitive tests is based on time or counts, which may have skewed distributions, it might be inappropriate to treat data from such tests as continuous indicators in CFA. Further, even for scores with less skewed distributions, the apparent gain of information from the continuous indicators approach comes at the expense of the assumption of linearity. When that assumption has been tested with cognitive test data, it has been found to be violated [e.g., 28, 30, 31].

*Variables*

*Neuropsychological Test Batteries.* Batteries of neuropsychological tests were administered in each study to measure a variety of cognitive functions. Each study administered at least two and as many as 19 cognitive tests. Many tests included multiple indicators (e.g., forwards and backwards subtests of the Digit Span test). We inventoried available cognitive tests in each study and identified tests in common across studies (table 2). We evaluated the internal consistency of each study's battery using Cronbach's α [32]. We used parallel analysis with screen plots to characterize the number of factors underlying each battery [33].

*Average Cortical Thickness.* MRI scans were conducted within one month of testing at each ADNI study visit. MRI data acquisition in ADNI is documented elsewhere (http://www.adni-info.org/Scientists/MRIProtocols.aspx). Briefly, 1.5 Tesla scanners were used to collect high-resolution sagittal three-dimensional T1-weighted Magnetization Prepared RApid Gradient Echo (MP-RAGE) scans with voxel sizes of 1.1 × 1.1 × 1.2 millimeters. To maximize reliability, MP-RAGE sequences were optimized for each study scanner. Scanners were routinely quality checked using a phantom by investigators at the Mayo site [19, 34]. Cortical thickness measurements were determined from MP-RAGE scans and processed through the longitudinal Freesurfer pipeline (version 5.1) [35, 36].

*Hippocampal Volume.* Hippocampal volume data were obtained as previously described [37].

*Statistical Analysis*

*Single Common Test Approach.* The total scores for the MMSE were available in all studies. The MMSE is a 30-point cognitive test of global mental status comprised of questions spanning different cognitive domains [38]. To facilitate direct comparisons with other approaches, we rescaled the MMSE in the pooled sample to have a mean of 50 and SD of 10 from the first study visit at which AD was diagnosed.

*Standardize and Average Approach.* We standardized each neuropsychological test to the mean and standard deviation (SD) from the first study visit in the sample, which was the earliest visit with a diagnosis of AD. Thus, for any test within a study, a score of 0 reflects the average of all scores from first visits of people with AD, and scores of +1 and –1 reflect one standard deviation above and below that average also for first visits of people with AD. We then averaged performance across available tests.

*CFA with Continuous Indicators Approach.* We obtained longitudinal latent cognitive ability scores for each observation across studies using CFA. Tests or subtests in common across studies serve to anchor the metric across studies [9, 10, 28]. We used maximum likelihood estimation with robust variance estimation in M*plus* (version 7.11, Muthen & Muthen, Los Angeles Calif., 1998–2008) to estimate the models. The model provides factor scores equivalent to those from a model with individual factors at each time point that more explicitly models longitudinal change.

Neuroepidemiology 2014;43:194–205
DOI: 10.1159/000367970



**Table 2.** Neuropsychological tests in each study

| Test name | Variables from the tests | Aging, demographics and memory study (ADAMS) | AD neuroimaging initiative (ADNI) | Rush memory and aging project (MAP) | Religious orders study (ROS) | National Alzheimer's coordinating center (NACC) | Cache county study on memory health and aging (Cache) | Myriad Tarenflurbil phase III clinical trial | Lilly Semagacestat phase III clinical trial | Add Neuro Med | Adult change in thought (ACT) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mini mental status exam (MMSE) | sum score | × | × | × | × | × | × | × | × | × | ×* |
| Boston naming test | 15-item, 30-item, 60-item | 15-item | 30-item | 15-item | 15-item | 30-item | 30-item | | | 15-item | 15-item |
| Semantic fluency | animals (A), vegetables (V) | A | A, V | A | A | A, V | A | A | | A | A |
| Digit span test | forward and backwards | × | × | × | × | × | × | × | | | |
| Logical memory I and II, Wechsler memory scale | story A, immediate and delayed recall; story B, immediate and delayed recall | × | × | × | × | × | | | | | × |
| Trail making test | parts A and B | × | × | | | × | × | | | | × |
| Word list learning (CERAD battery) | immediate and delayed recall | × | | × | × | | × | | | × | |
| Symbol-digit modalities test | number of number/symbol matches | × | | × | × | | | | | | |
| Controlled oral word association test | sum of F,A,S words | × | | | | | × | × | | | |
| Constructional praxis | total, delay, recognition | × | | | | | × | | | × | × |
| Digit symbol substitution, Wechsler memory scale-revised | number of number/symbol matches | | × | | | × | | | | | |
| ADAS-Cog | sum score | | × | | | | | × | × | × | |
| Auditory verbal learning test | trial 1-5 sum of recall, delayed recall | | × | | | | | × | | | |
| Line orientation | total correct | | | × | × | | | | | | |
| Number comparisons | total correct | | | × | × | | | | | | |
| Digit ordering | total correct | | | × | × | | | | | | |
| East Boston story test | immediate and delayed recall | | | × | × | | | | | | |
| Ravens progressive matrices | total correct | | | × | × | | | | | | |
| Alphabet span | longest span | | | | × | | | | | | |
| Hopkins verbal learning test | trial 1–3 sum of recall, delayed recall | | | | | | × | | | | |
| Shipley | number correct | | | | | | × | | | | |
| Paired associates | easy and hard trials; immediate and delayed | | | | | | | × | | | × |
| Digit cancellation | number correct | | | | | | | × | | | |
| Object recognition | total recall | | | | | | | | | × | |
| Word list learning | recall and recognition | | | | | | | | | | × |
| WAIS Information | total recall | | | | | | | | | | × |
| WAIS comprehension | total recall | | | | | | | | | | × |
| Mattis dementia rating scale | total score | | | | | | | | | | × |
| Total number of cognitive tests | | 10 | 9 | 12 | 13 | 7 | 10 | 8 | 2 | 7 | 11 |
| Total number of cognitive test indicators | | 16 | 13 | 18 | 19 | 11 | 12 | 12 | 2 | 7 | 16 |
| Cronbach's alpha by dataset | | 0.91 | 0.90 | 0.94 | 0.95 | 0.89 | 0.84 | 0.89 | 0.67 | 0.85 | 0.84 |

* The MMSE in ACT was calculated using items from the Cognitive Abilities Screening Instrument (CASI).

*CFA with Categorical Indicators Approach.* Prior to being used as indicators in a CFA model, we categorized each cognitive test score, using identical cutoffs across studies (online suppl. table 1; for all online suppl. material, see www.karger.com/doi/10.1159/000367970). We used an equal interval approach to categorization to preserve the distribution of the original test. As in the continuous indicator CFA approach, tests or sub-tests in common serve to anchor the metric across studies and we used a maximum likelihood estimator with robust standard error estimation in M*plus*. The model is consistent with an item response theory graded response model [39–41].

*External Scaling of the Factor Scores for Stability.* Using methods described in detail elsewhere [30], we externally scaled factors from the continuous and categorical indicator CFA models so that a mean of 50 and SD of 10 represented older adults aged 70 years and older in the United States by fixing model parameters in the pooled data to their counterparts from a CFA from the Aging, Demographics and Memory Study (ADAMS) [42].

*Missing Data Handling.* The common test and standardize and average approaches use a complete case analysis, which assumes data are missing completely at random. The CFA approaches make less restrictive assumptions about missing data by assuming the missing nature of data in specific cognitive tests are missing at random conditional on variables in the measurement model. This is handled using maximum likelihood methods, and is a reasonable approach for measuring the general cognitive performance because an implicit assumption is that tests are exchangeable with each other.

*Simulation to Demonstrate Comparability of Summary Scores Across Datasets.* To demonstrate that derived scores from the standardize and average approach, CFA with continuous indicators and CFA with categorical indicators were comparable across different studies that administered different sets of cognitive tests, we conducted Monte Carlo simulations. Based on empirical correlations among cognitive tests, we simulated 100,001 observations with complete cognitive data. We then calculated summary scores based on each of the approaches for each observation using tests from each study. We examined bias and precision in test-specific cognitive scores with respect to the true score (whether an average of standardized values, CFA of continuous items, or CFA of categorical items) that used all available items using Bland-Altman plots [43]. Simulation is not needed to evaluate comparability of the MMSE because no equating was done on that measure.

*Comparison of Measurement Approaches.* We compared the approaches in three sets of analyses. First, we correlated the measures using baseline data in the pooled sample. Second, we modeled the annual rate of change using the random effects models to compare the relative magnitudes of change detected by the approaches [44]. The timescale was the time from the earliest onset of AD symptoms. We calculated the sample size needed to detect a 25% annual decline in cognitive performance with 80% power using each approach. We included terms for age, sex, and years of education in these models. We selected a magnitude of 25% because this is a common effect size in other genetic studies. We determined sample size using this equation:

$$(2*SD\_CHANGE^2*(1.96 + 0.84)^2) / (EFFECT\_SIZE*MEAN\_CHANGE)^2 \quad (1)$$

There was a modest amount of missing data for demographic variables; so we used multiple imputation procedures with 22 random draws to account for this [45]. Third, using up to six years of data from ADNI, we examined associations of change in average cortical thickness and hippocampal volume with change in cognitive performance, using joint process growth curve models [46].

## Results

The full sample included 10,875 older adults with AD (table 1). Longitudinal cognitive data were available from all studies, with up to 17 measurement occasions (median 4) spanning up to 19 years (median 1.6 years). Participants were on average 77 years old (range 60, 110). The sample was 55% female, 86% white, and well educated: 41% had a college education and 20% had a graduate level of education or higher.

### *Neuropsychological Test Batteries*

We identified 60 indicators from 28 cognitive tests administered across all studies (table 2). Cronbach's α estimates were above 0.84 for each study except for the Semagacestat trial (α = 0.67), for which only two tests were administered (MMSE and ADAS-Cog) (table 2). The MMSE was administered in each study. Other common tests included Digit Span (6 studies), Trail Making Test (4 studies), ADAS-Cog (4 studies), Boston Naming Test (7 studies), and Logical Memory (5 studies) (table 2). Parallel analysis with scree plots suggested that unidimensionality was sufficiently met in each study.

### *Simulation to Demonstrate Comparability of Summary Scores Across Datasets*

Bland-Altman plots are provided in online supplemental figures 1–3. For the CFA with categorical indicators approach (online suppl. fig. 1), we found high correlations between study-specific factors and known true factors (r's >0.90), minimal bias (e.g., in ADNI, 0.21 points, or a 0.021 standard deviation difference), and lack of systematic deviation over the range of cognitive functioning. An exception is the Semagacestat study, for which only two cognitive tests were administered. Scores from that dataset showed minimal bias (0.31 points), but less precision (r = 0.90). These characteristics suggest that factor scores in each dataset derived using CFA with categorical indicators are on the same metric. Although we did not specify a priori an acceptable range of bias that would cause concern, we note that the magnitude of the bias from our simulations is approximately 10 times smaller than the observed annual rate of change in the CFA with categorical indicators approach (2.57 points, or 0.257 SD units) (table 3).

**Table 3.** Annual rates of cognitive decline among persons with AD using different cognitive summary scores (n = 10,875)

| Summary score approach | Average pace of cognitive decline (per 10 SD units) | Standard deviation of the random slope | Sample size needed to detect 25% decline with 80% power* |
|---|---|---|---|
| Common item (MMSE) | –3.38 | 3.56 | 277 |
| Standardize and average | –3.10 | 3.34 | 291 |
| Continuous indicator CFA | –2.96 | 3.32 | 315 |
| Categorical indicator CFA | –2.57 | 2.47 | 232 |

Estimated means and standard errors were calculated from random effects models of each cognitive predictor regressed on time since a dementia diagnosis. Each summary cognitive score was standardized to a have a mean of 50 and SD of 10. Models were adjusted for age, sex, and years of education.

* We selected a magnitude of 25% because this is a common effect size in other genetic studies. To determine the sample size needed to detect a smaller effect (e.g., 3%), see equation 1 in the Methods section.

SD = Standard deviation; MMSE = Mini-Mental State Examination; CFA = confirmatory factor analysis.

Findings from the simulation were similar for the continuous indicator CFA approach (online suppl. fig. 2), although the spread of the differences between true and study-specific estimates were 1.5 to 2 times wider than those for the categorical indicator CFA approach. Findings from the simulation regarding the standardize and average approach (online suppl. fig. 3) yielded very different results that suggested large amounts of bias of up to 11 points (1.1 SD units) in some studies that were more extreme especially at the lower ends of the spectrum. The spread of the differences between true and study-specific estimates were comparable to the spread for the continuous indicator CFA approach.

Taken together, these results from simulation analyses suggest that CFA with categorical indicators approach provided less bias and more precision than other approaches. Although the CFA with continuous indicators approach demonstrated minimal bias, precision was worse than the CFA with categorical indicators approach. The standardize and average approach resulted in bias and low precision.

*Relationships between Approaches*

Figure 2 shows score distributions and correlations among the approaches. Correlations were all above 0.84. The MMSE had the lowest correlations with other approaches, and was highly left-skewed. The standardize and average approach and CFA scores appeared to have fairly normal distributions (fig. 2). The two CFA approaches were highly correlated (r = 0.94). In the scatterplots between the categorical indicator CFA and other approaches (bottom row), overall bowing in the scatterplots and heteroskedasticity in the more cognitively impaired range suggests nonlinearity in the component tests, which may not satisfy strict assumptions of linearity made in the CFA model with continuous indicators. Further, diverging prongs in scatterplots against the standardize and average approach (second column) suggests that this approach does not provide scores on the same metric across datasets.

*Power to Detect Change*

All approaches yielded comparable rates of cognitive decline, between 0.026 and 0.034 SD units per year, after scaling them to have a mean of 50 and SD of 10 (table 3). While point estimates for change were comparable, model-estimated standard deviations were the largest for the single common test approach and smallest for the categorical indicator CFA (table 3). These characteristics are also reflected in the power analyses because the minimum sample size required to have 80% power to detect a 25% cognitive decline was the smallest for the categorical indicator CFA approach (n = 232) (table 3).

*Criterion Validity Using Imaging Markers*

We examined criterion validity against average cortical thickness and hippocampal volume changes using up to six years of longitudinal data from ADNI. Changes in cognitive performance using each approach were strongly associated with changes in both mean average cortical thickness and hippocampal volume (table 4). The CFA with categorical indicators, however, provided the strongest standardized effects (Z = 7.0 for average cortical thickness; Z = 5.2 for hippocampal volume), which were calculated as the quotient of the estimated covariance and its standard error. Although differences in effects between cognitive measures were not statistically significant as indicated by comparisons in the lower part of ta-

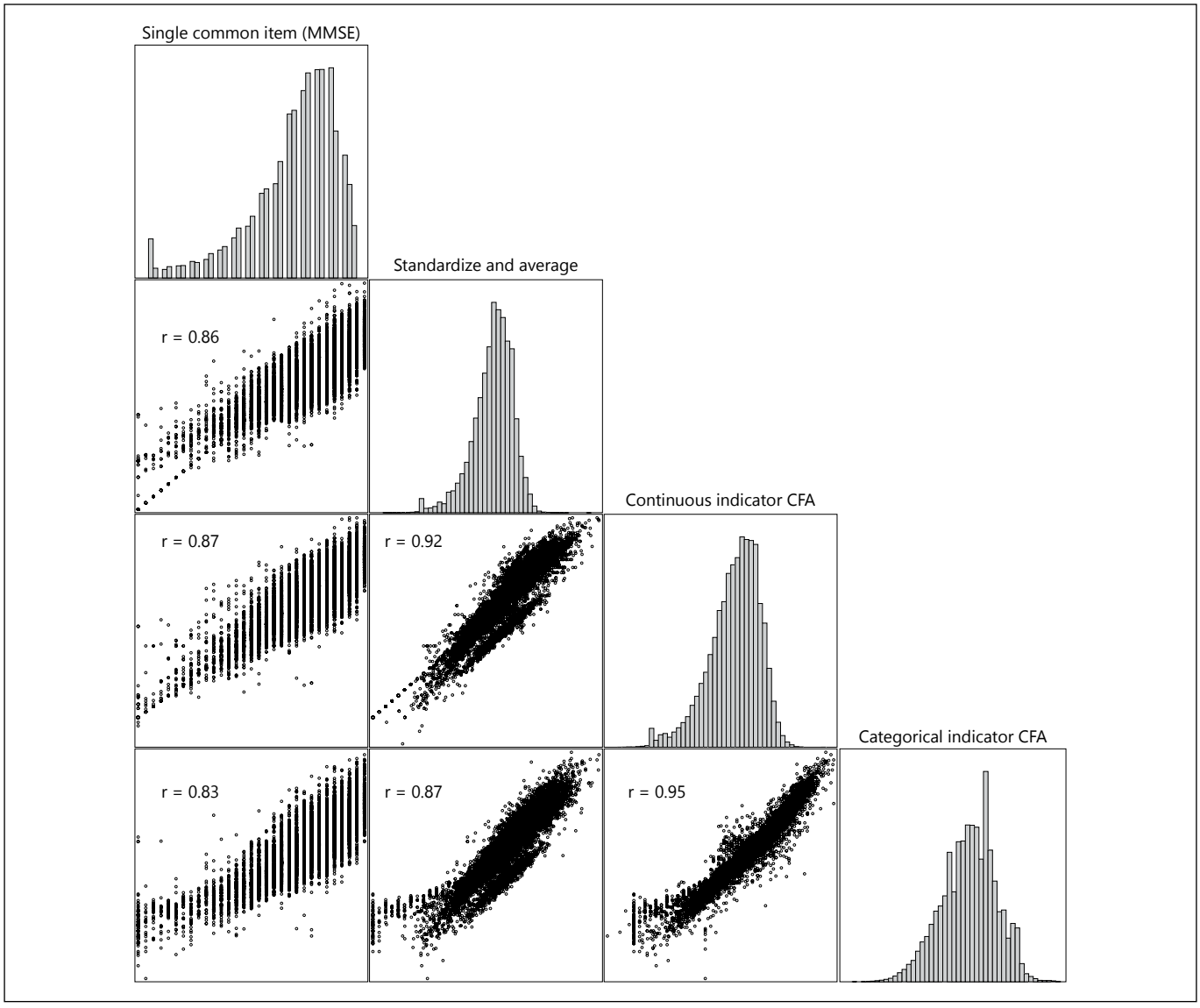


**Fig. 2.** Distributions and correlations for each summary cognitive score (n = 10,875). The on-diagonal figures show histograms for each approach in the pooled data. The off-diagonal figures show scatterplots of each approach against another. MMSE = Mini-Mental State Examination; GCP = general cognitive performance.

ble 4, corresponding standardized effects were lower using the common test (Z = 6.2; Z = 4.5), standardize and average (Z = 5.5; Z = 3.7), and CFA with continuous indicators (Z = 6.1; Z = 4.1) approaches.

## Discussion

We evaluated the four approaches for calibrating longitudinal cognitive performance among people with AD across nine studies with different but overlapping sets of neuropsychological tests. Factor analysis with categorical versions of test indicators produced the most precise estimates of the rate of change, reflected by improved power to detect differences in the rate of decline, and also had a stronger strength of association with changes in neuroimaging markers from ADNI. Missing data assumptions required for CFA with categorical indicators are no more restrictive than those of the other approaches we tested. This approach has broad applicability to directly compare cognitive performance in existing and future studies, making it a good choice for

**Table 4.** Associations between changes in each cognitive summary score and selected biomarkers: Results from ADNI (n = 431)

| Biomarker | (A) Common item (MMSE) | | (B) Standardize and average | | (C) Continuous indicator factor analysis | | (D) Categorical indicator factor analysis | |
|---|---|---|---|---|---|---|---|---|
| | Covariance of change (SE) | standardized estimate (z) | Covariance of change (SE) | standardized estimate (z) | Covariance of change (SE) | standardized estimate (z) | Covariance of change (SE) | standardized estimate (z) |
| Average cortical thickness | 0.08±0.01 | 6.15 | 0.04±0.01 | 5.50 | 0.06 (0.01) | 6.11 | 0.04 (0.01) | 7.00 |
| Hippocampal volume | 1.84±0.41 | 4.46 | 1.02±0.28 | 3.72 | 1.31 (0.32) | 4.09 | 0.96 (0.19) | 5.21 |
| Difference in z (p value) | | | | | | | | |
| Average cortical thickness | | | | | | | | |
| Difference with: | | | | | | | | |
| Standardize and average | | 0.65 (0.51) | | | | | | |
| Continuous indicator factor analysis | | 0.04 (0.97) | | −0.61 (0.54) | | | | |
| Categorical indicator factor analysis | | −0.85 (0.40) | | −1.50 (0.13) | | −0.89 (0.37) | | |
| Hippocampal volume | | | | | | | | |
| Difference with: | | | | | | | | |
| Standardize and average | | 0.74 (0.46) | | | | | | |
| Continuous indicator factor analysis | | 0.37 (0.71) | | −0.37 (0.71) | | | | |
| Categorical indicator factor analysis | | −0.75 (0.45) | | −1.49 (0.14) | | −1.12 (0.26) | | |

Associations between changes in selected biomarkers and rate of change in cognitive performance measured by four approaches to harmonization. Standardized estimates (z-scores) were calculated as the quotient of the estimated covariance and its standard error. Pairwise comparisons in the lower part of the table show z-scores and p-values for the difference in standardized estimates between each approach; none were statistically significantly different from each other at the $\alpha < 0.05$ level. MMSE = Mini-Mental State Examination; SE = standard error.

deriving a phenotype for genetic analyses of cognitive decline.

We observed that the CFA with categorical indicators required the lowest sample size to detect a 25% decline with 80% power. Theoretically, using more tests should be less susceptible to random fluctuations and thus yield more precise estimates of cognitive performance. Although the MMSE produced the next best sample size, it is notoriously unreliable in community-living samples [28, 47]. Its precision in our study may be attributable to the fact that we restricted the analysis to participants with AD. MMSE scores in the more impaired range where people with AD performance tend to be more precise than at less impaired levels.

One of the more surprising results of our analyses is the poor performance of the very commonly used standardize and average approach. Simulations suggested considerable bias in each study, which is confirmed by scatterplots of observed data (fig. 2). In simulations, the range of differences between true and study-specific estimates was between 1.5 and 2 times larger than that for the CFA with categorical indicators approach, suggesting imprecision. These results strongly suggest that the field should reconsider the ubiquity with which this approach is employed. While it is always feasible to obtain a z-score for any particular set of tests, and average those z-scores for any battery of tests, the only advantage of that approach is that it is feasible. In the 21st century, widespread use of computational infrastructure permits us to use better approaches to the problem. All of our analyses suggest that the categorical indicator CFA approach is superior to the standardize and average approach.

In addition to the empirical challenges of the standardize and average approach to calibrating cognitive performance across datasets, it is not theoretically sound for two reasons. First, different combinations of cognitive tests, equally weighted within a study to construct the summary score, could reflect qualitatively different constructs across studies. For example, one study may include two tests of memory and four tests of attention, while another study might include four tests of memory, one test of attention, and one test of visuospatial ability. Both studies administered six tests, but the second study's composite score is more heavily laden with memory tests. Failure to address unequal weighting of the different cognitive tests can bias estimates of change in cognitive performance. This imbalance is likely responsible for the diverging prongs present in scatterplots of the standardize and average approach compared with other approaches in figure 2. Conceptual differences across studies are even larger if the number of tests administered and the proportions of cognitive domains represented vary. This is

precisely the situation we face with the data available to us (table 2). Using standardized and averaged scores makes sense within a single study or when precisely the same measures are available across studies. Factor analysis approaches address these concerns by allowing cognitive tests to be weighted differently within a study and also by accommodating over-represented cognitive domains using a bifactor approach as we did for memory tests. A second reason why the standardize and average approach is not theoretically sound for cognitive change among persons with AD is that it assumes that test difficulty is evenly distributed over the latent cognitive ability trait. This is clearly not the case for individual tests such as the MMSE [28, 47, 48], and likely not true for the combinations of multiple tests in a neuropsychological battery. The categorical indicator CFA approach does not make this assumption.

Although our goal was to identify a high-quality approach to use to generate inter-study genetic phenotypes for longitudinal cognitive performance in AD, this work may have applicability for other applications in epidemiology and cognitive aging. The need to calibrate previously collected cognitive phenotypes across studies is a well-recognized challenge in neuropsychology and cognitive aging [49, 50]. Over 500 cognitive tests are available for use epidemiologically and clinically, complicating efforts to synthesize information [12]. High-quality harmonization requires maximizing comparability and precision of phenotypes across datasets and minimizing misclassification [5]. Harmonization is not necessary to address a scientific question if a single study can be used. However, the sample size from any single study may be too small to draw reliable conclusions, especially in genetic studies where needed sample sizes may be large. Properties of the categorical indicator CFA approach make it an appropriate method for calibrating longitudinal cognitive performance across studies and batteries as long as some tests overlap.

Strengths of this study include the large number of prospectively collected datasets with a rich diversity of sample characteristics and sampling strategies. This study was possible through collaborative efforts with investigators from studies across multiple institutions willing to share their data with us. Further, our approach is scalable with more data. The quality of the links that hold datasets together improve as more data are added. Methods for obtaining a general cognitive factor based in generalized item response models are robust to the inclusion of other neuropsychological tests from different batteries and so more data should only strengthen the approach. For example, the approach could be used to take advantage of the wealth of historical data from Alzheimer's Disease Research Centers before the development of the UDS battery.

An important limitation is that calibrating cognitive performance across studies using CFA relies on cognitive tests in common across datasets. Although one simulation reported that at least five items in a calibration exercise such as ours is enough to provide a link [51], at least one previous study used just a single indicator to combine samples [52]. All datasets had at least one indicator in common (MMSE), and eight datasets had at least six test indicators in common (table 2). Another limitation is that the global measure of cognitive performance we evaluated is not a substitute for individual cognitive domains such as memory or executive functioning. A final limitation is that, while the categorical indicator CFA approach is scalable to other datasets, when conducting integrative data analysis, it is important to pay attention to the sampling frame and study designs. We restricted our sample to persons with AD because that was relevant to a scientific question. Other studies should carefully review exclusion and inclusion criteria, as well as retention procedures, for each study in a pooled sample.

Future developments in AD research are likely to come from research in biology and genetics and 'big data' approaches [7]. Availability of genome-wide data techniques has created the need for large sample sizes to detect genetic associations. Calibration of cognitive performance across studies can contribute to physiological research that may reveal biological mechanisms to help researchers determine the etiology underlying heterogeneity in rates of clinical progression of AD. This knowledge may be translated into novel therapeutic targets. Our results suggest that the CFA with categorical indicators approach produces operational phenotypes with greater precision than the other approaches we considered, enabling us to better define the cognitive phenotypes used in genetic studies that combine multiple datasets.


## Funding

This study was supported by grants from the National Institutes of Health: R03 AG045494 (PI: Gross), R01 HG002213 and K24 AG027841 (PI: Green), U01 AG032984 (Alzheimer's Disease Genetics Consortium). Drs. Sherva and Green were supported by a grant from the Alzheimer's Association. Dr. Crane was supported by NIA U01 AG06781 and NHGRI U01 HG006375 (Eric Larson, PI).

The National Alzheimer's Coordinating Center (NACC) database is supported by U01 AG016976. The Adult Changes in Thought (ACT) study was supported by NIA U01 AG06781 (Larson). The Cache County Study was supported by the Alzheimer's Association (MNIRG-11–205368), the BYU Gerontology Program and the NIH (R01 AG11380, R01 AG021136, P30 NS069329–01, R01 AG042611). ROS/MAP was supported by NIH grants P30AG10161, R01AG17917, and R01AG17917.

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; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; 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; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of 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 (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.

No authors claim financial conflicts of interest. The contents do not necessarily represent the views of the funding entities. Funders had no deciding roles in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

## Corporate Authorship List for GENAROAD Consortium

*Albert Einstein College of Medicine:* Charles Hall; Richard Lipton.

*Boston University School of Medicine:* Lindsay Farrer; Richard Sherva.

*Brigham and Women's Hospital/Harvard Medical School:* Lori B. Chibnik; Phillip De Jager; Robert C. Green; Leo B. Waterston.

*Brigham Young University:* John Kauwe.

*Cardiff University:* Melanie Dunstan; Julie Williams.

*Department of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland:* Hilkka Soininen.

*Eli Lilly and Company:* Leanne M. Munsie.

*Harvard School of Public Health:* Peter Kraft.

*Indiana University School of Medicine:* Andrew Saykin,

*INSERM U 558, University of Toulouse, Toulouse, France:* Bruno Vellas.

*Institute of Gerontology and Geriatrics, University of Perugia, Perugia, Italy:* Patrizia Mecocci.

*Johns Hopkins Bloomberg School of Public Health:* Alden L. Gross.

*King's College London:* Chantal Bazenet; Richard Dobson; Simon Lovestone; Stephen Newhouse.

*Lille 2 University Hospital/Inserm/Institut Pasteur de Lille/Bordeaux 2 University/Isped:* Philippe Amouyel; Céline Bellenguez; Carole Dufouil.

*Medical University of Lodz, Lodz, Poland:* Iwona Kloszewska.

*Pfizer:* Jens Wendland; Ashley Winslow.

*Rush University Medical Center:* David A. Bennett.

*Third Department of Neurology, Aristotle University, Thessaloniki, Greece:* Magda Tsolaki.

*University of Pennsylvania:* Gerard Schellenberg.

*University of Washington:* Paul K. Crane; Shubhabrata Mukherjee.

*Utah State University:* Christopher Corcoran; Ronald Munger; Maria Norton; JoAnn Tschanz.

*Washington University in St. Louis:* Carlos Cruchaga.

## References

1 Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, et al; GENEVA Consortium: The Gene, Environment Association Studies Consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 2010; 34:364–372.

2 Manolio TA, Rodriguez LL, Brooks L, Abecasis G; Collaborative Association Study of Psoriasis, Ballinger D, et al: New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 2007;39:1045–1051.

3 Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, et al; CHARGE Consortium: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2009;2:73–80.

4 Sherva R, Tripodis Y, Bennett DA, Chibnik LB, Crane PK, de Jager PL, et al; Alzheimer's Disease Genetics Consortium: Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. Alzheimer's Dement 2014;10:45–52.

5 Seminara D, Khoury MJ, O'Brien TR, Manolio T, Gwinn ML, Little J, et al; Human Genome Epidemiology Network; Network of Investigator Networks. The emergence of networks in human genome epidemiology: challenges and opportunities. Epidemiology 2007;18:1–8.

6 Zeggini E, Ioannidis JP: Meta-analysis in genome-wide association studies. Pharmacogenomics 2009;10:191–201.

7 Khachaturian AS, et al: Big data, aging, and dementia: pathways for international harmonization on data sharing. Alzheimers Dement 2013;9:S61–S62.

8 Alzheimer's Association Expert Advisory Workgroup on NAPA: Workgroup on NAPA's scientific agenda for a national initiative on Alzheimer's disease. Alzheimers Dement 2012;8:357–371.

9 Bauer DJ, Hussong AM: Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. Psychol Methods 2009;14: 101–125.

10 Curran PJ, Hussong AM: Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods 2009;14:81–100.

11 Barnes DE, Cenzer IS, Yaffe K, Ritchie CS, Lee SJ; Alzheimer's Disease Neuroimaging Initiative: A point-based tool to predict conversion from mild cognitive impairment to probable Alzheimer's disease. Alzheimer's Dement 2014, DOI: 10.1016/j.jalz.2013.12.014, Epub ahead of print.

12 Lezak MD, Howieson DB, Loring DW: Neuropsychological Assessment. New York, Oxford University Press, 2004.

13 Akshoomoff N, Newman E, Thompson WK, McCabe C, Bloss CS, Chang L, et al: The NIH Toolbox Cognition Battery: results from a large normative developmental sample (PING). Neuropsychology 2014;28:1–10.

14 Weintraub S, Salmon D, Mercaldo N, Ferris S, Graff-Radford NR, Chui H, Cummings J, DeCarli C, Foster NL, Galasko D, Peskind E, Dietrich W, Beekly DL, Kukull WA, Morris JC: The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery. Alzheimer Dis Assoc Disord 2009;23:91–101.

15 den Heijer T, Geerlings MI, Hoebeek FE, Hofman A, Koudstaal PJ, Breteler MM: Use of hippocampal and amygdalar volumes on magnetic resonance imaging to predict dementia in cognitively intact elderly people. Arch Gen Psychiatry 2006;63:57–62.

16 den Heijer T, van der Lijn F, Koudstaal PJ, Hofman A, van der Lugt A, Krestin GP, et al: A 10-year follow-up of hippocampal volume on magnetic resonance imaging in early dementia and cognitive decline. Brain 2010;133:1163–1172.

17 Sluimer JD, van der Flier WM, Karas GB, Fox NC, Scheltens P, Barkhof F, Vrenken H: Whole-brain atrophy rate and cognitive decline: longitudinal MR study of memory clinic patients. Radiology 2008;248:590–598.

18 Sluimer JD, Bouwman FH, Vrenken H, Blankenstein MA, Barkhof F, van der Flier WM, Scheltens P: Whole-brain atrophy rate and CSF biomarker levels in MCI and AD: a longitudinal study. Neurobiol Aging 2010;31:758–764.

19 Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al; Alzheimer's Disease Neuroimaging Initiative: The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement 2013;9:e111–e194.

20 Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS: Overview and findings from the rush memory and aging project. Curr Alzheimer Res 2012;9:646–663.

21 Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS: Overview and findings from the religious orders study. Curr Alzheimer Res 2012;9:628–645.

22 Breitner JC, Wyse BW, Anthony JC, Welsh-Bohmer KA, Steffens DC, Norton MC, et al: APOE-epsilon4 count predicts age when prevalence of AD increases, then declines: the Cache County Study. Neurology 1999;53:321–331.

23 Green RC, Schneider LS, Amato DA, Beelen AP, Wilcock G, Swabb EA, et al; Tarenflurbil Phase 3 Study Group: Effect of tarenflurbil on cognitive decline and activities of daily living in patients with mild Alzheimer disease: a randomized controlled trial. JAMA 2009;302:2557–2564.

24 Doody RS, Raman R, Farlow M, Iwatsubo T, Vellas B, Joffe S, et al; Semagacestat Study Group: A phase 3 trial of semagacestat for treatment of Alzheimer's disease. N Engl J Med 2013;369:341–350.

25 Lovestone S, Francis P, Kloszewska I, Mecocci P, Simmons A, et al: AddNeuroMed – the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Ann N Y Acad Sci 2009;1180:36–46.

26 Kukull WA, Higdon R, Bowen JD, et al: Dementia and Alzheimer disease incidence: a prospective cohort study. Arch Neurol 2002;59:1737–1746.

27 Morris JC, Weintraub S, Chui HC, Cummings J, Decarli C, et al: The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. Alzheimer Dis Assoc Disord 2006;20:210–216.

28 Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al: Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol 2008;61:1018–1027.

29 Folstein MF, Folstein SE, McHugh PR: 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–198.

30 Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK: Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology 2014;42:144–153.

31 Proust-Lima C, Amieva H, Dartigues JF, Jacqmin-Gadda H: Sensitivity of four psychometric tests to measure cognitive changes in brain aging-population-based studies. Am J Epidemiol 2007;165:344–350.

32 Cronbach LJ: Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

33 Hayton JC, Allen DG, Scarpello V: Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. Organ Res Methods 2004;7:191.

34 Kruggel F, Turner J, Muftuler LT; Alzheimer's Disease Neuroimaging Initiative: Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. Neuroimage 2010;49:2123–2133.

35 Dale AM, Fischl B, Sereno MI: Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 1999;9:179–194.

36 Prabhakaran V, Nair VA, Austin BP, La C, Gallagher TA, Wu Y, et al: Current status and future perspectives of magnetic resonance high-field imaging: a summary. Neuroimaging Clin N Am 2012;22:373–397.

37 Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31:968–980.

38 Feher EP, Mahurin RK, Doody RS, Cooke N, Sims J, Pirozzolo FJ: Establishing the limits of the Mini-Mental State. Examination of 'subtests'. Arch Neurol 1992;49:87–92.

39 Lord FM: The relation of test score to the trait underlying the test. Educational and Psychological Measurement 1953;13:517–549.

40 Samejima F: Estimation of latent ability using a response pattern of graded scores. Psychometrika 1969;35:139.

41 Takane Y, de Leeuw J: On the relationship between item response theory and factor analysis of discretized variables. Psychometrika 1987;52:393–408.

42 Langa KM, Plassman BL, Wallace RB, Herzog AR, Heeringa SG, Ofstedal MB, et al: The Aging, Demographics, and Memory Study: study design and methods. Neuroepidemiology 2005;25:181–191.

43 Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310.

44 Laird NM, Ware JH: Random effects models for longitudinal data: an overview of recent results. Biometrics 1982;38:963–974.

45 Rubin DB: Multiple Imputation for Nonresponse in Surveys. New York, NY, Wiley & Sons, 1987.

46 Muthén B: Latent variable hybrids: overview of old and new models; in Hancock GR, Samuelsen KM (eds): Advances in latent variable mixture models. Charlotte, NC, Information Age Publishing. 2008, pp 1–24.

47 Mungas D, Reed BR: Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. Stat Med 2000;19:1631–1644.

48 Wouters H, van Gool WA, Schmand B, Zwinderman AH, Lindeboom R: Three sides of the same coin: measuring global cognitive impairment with the MMSE, ADAS-cog and CAMCOG. Int J Geriatr Psychiatry 2010;25:770–779.

49 Bennett SN, Caporaso N, Fitzpatrick AL, Agrawal A, Barnes K, Boyd HA, et al; GENEVA Consortium: Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. Genet Epidemiol 2011;35:159–173.

50 Hendrie HC, Albert MS, Butters MA, Gao S, Knopman DS, Launer LJ, et al: The NIH Cognitive and Emotional Health Project. Report of the Critical Evaluation Study Committee. Alzheimer's Dement 2006;2:12–32.

51 Wang W-C: Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. Journal of Experimental Education 2004;72:221–261.

52 Jones RN, Fonda SJ: Use of an IRT-based latent variable model to link different forms of the CES-D from the Health and Retirement Study. Soc Psychiatry Psychiatr Epidemiol 2004;39:828–835.