Agreement in Cognitive and Clinical Assessments in Alzheimer's Disease

Rochelle E. Tractenberg^{a, b} Paul S. Aisen^b

^aDepartments of Neurology, Biostatistics, Bioinformatics and Biomathematics, and Psychiatry, Georgetown University School of Medicine, Washington, D.C., and ^bDepartment of Neurology, University of California, San Diego, La Jolla, Calif., USA

Key Words

Alzheimer's disease • Dementia • Outcomes assessment methods

Abstract

Aims: To estimate agreement among scores on three common assessments of cognitive function. **Method:** Baseline responses on the Alzheimer's Disease Assessment Scale – Cognitive, Clinical Dementia Rating, and the Mini-Mental State Examination were obtained from two clinical trials ($n = 138$ and $n = 351$). A graphical method of examining agreement, the means-difference or Bland-Altman plot, was followed by Levene's test of the equality of variance corrected for multiple comparison within each sample. **Results:** 70–78% of variability was shared by one factor, suggesting that all three instruments reflect cognitive impairment. However, agreement among tests was significantly worse for individuals with greater-than-average, relative to individuals with less-than-average, cognitive impairment. **Conclusions:** Worse agreement between tests, as a function of increasing cognitive impairment, implies that interpretation of these tests and selection of coprimary cognitive impairment outcomes may depend on impairment level.

Copyright © 2009 S. Karger AG, Basel

Introduction

Cognitive function and/or impairment are the most widely assessed [1–6] symptoms in studies of cognitive aging and Alzheimer's disease (AD). Importantly, there is a fundamental difference between the more symptom-specific assessment of 'cognitive function', where the presence of particular cognitive impairments is determined, and a more global 'clinical' assessment, which emphasizes overall functioning but depends on cognitive functioning to some extent. For example, one of each of these types of assessments (cognitive, global) is mandated in clinical efficacy trials for AD interventions by the Food and Drug Administration [7, p. 24]. The present analysis focuses on agreement in estimates of cognitive function, a latent (not directly observable) construct.

Association describes a general linear relationship between two variables and the strength of this association is estimated in the calculation of a correlation coefficient [8–10]. In 1983, Altman and Bland [11, see also 12] described a method for comparing a pair of methods for measurement of continuous outcomes. The method involves examining the difference between two measurements relative to the average value of the two measurements as an indication of their *agreement*, rather than the *association* between two such measures [11, 13]. We used the Bland-Altman (BA) method and arguments in an examination of three assessments that are extremely common in diagnosing, staging, and monitoring cognitive function in cognitive aging as well as in AD clinical research. In clin-

ical trial contexts, these three tests are frequently used together in the same study: the Alzheimer's Disease Assessment Scale – Cognitive (ADAS-Cog) [5, 14], Clinical Dementia Rating (CDR) [3], and the Mini-Mental State Examination (MMSE) [6]. These instruments were developed over the 1970s to 1980s with different specific goals, but are currently employed and interpreted in the same general ways in persons with AD, namely as reflections of an individual's cognitive functioning. The Alzheimer's Disease Cooperative Study (ADCS) [15] recently completed two multicenter treatment trials for AD. Each study employed scores on the MMSE and the CDR to determine inclusion eligibility, and change in the ADAS and CDR to ascertain outcome. These datasets provided excellent opportunities to explore agreement and interchangeability of these tests and to replicate results in two independent cohorts.

We preceded our BA evaluations of agreement with a factor analysis seeking evidence of a similar underlying factor structure for these three instruments, i.e., that all reflect a single underlying construct. This is a critical feature to support treating results as 'exchangeable' across clinical settings; a single underlying factor is critical for any pair of assessments to meet the FDA guidelines of two outcomes (one specific, one global) [7, p. 24].

Materials and Methods

Methods

This is a secondary analysis of data collected at baseline in two NIH-funded multicenter placebo-controlled clinical trials evaluating the effects of anti-inflammatory drugs on cognitive decline in AD. Both studies were approved by the institutional review boards of all participating centers.

Prednisone Study

Aisen et al. [16] conducted a multicenter randomized placebo-controlled clinical trial to determine the effect of the steroid anti-inflammatory drug prednisone (PR) on cognitive decline in 138 persons who met National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer Disease and Related Disorders Association (NINCDS-ADRDA) [17] criteria for probable AD. Use of cholinesterase inhibitors was not allowed.

Nonsteroidal Study

Aisen et al. [18] conducted a multicenter randomized placebo-controlled clinical trial to determine the effect of the nonsteroidal (NS) anti-inflammatory drugs rofecoxib and naproxen on cognitive decline in 351 persons with NINCDS-ADRDA diagnoses of probable AD. At baseline, 68% of this sample was on a stable dose of a cholinesterase inhibitor.

Inclusion and exclusion criteria were similar for the two studies; cholinesterase inhibitor use and specific exclusions for the

presence of conditions that increased the risk of adverse events associated with the treatment in either study were different. Table 1 presents the descriptive statistics for the two cohorts.

Materials

Both cohorts were administered the same three assessments at baseline. The primary symptom-specific outcome measure in both trials was the ADAS-Cog [5, 14]. The ADAS-Cog, designed specifically for use with persons with AD, measures memory, attention, reasoning, language, orientation and praxis. Scores range from 0 to 70 with higher scores reflecting higher cognitive impairment.

The CDR [3, 4] was the global outcome measure in both studies. The CDR is based on a semistructured interview of both the patient and a knowledgeable informant. The instrument rates each of six domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Each domain ('box') is rated as normal (= 0) or as involving questionable (= 0.5), mild (= 1), moderate (= 2) or severe (= 3) impairment, from which a global rating of dementia severity, if present, is derived [4]. The CDR global rating (0–3) describes the severity of dementia, and in addition to serving as a global outcome measure, eligibility in both studies required a global CDR ≥ 0.5 . Rather than using the global rating, the box scores can also be summed (0–18), with higher box scores reflecting higher impairment (worse dementia) [Morris, pers. commun.]; we used the sum of box scores rather than the global CDR rating.

The MMSE [6] was part of the screening process for both studies. Full-scale MMSE scores are calculated as the sum of 1/0 correct/incorrect answers to a series of questions, so that scores range from 0 to 30 (worst to best). The MMSE was developed as a measure of overall cognitive function at bedside, but it is widely used to monitor or assess cognitive functioning (as opposed to dementia). We subtracted observed scores from 30 to derive an index of cognitive impairment comparable in direction to that of the other two instruments (i.e., higher = worse).

Statistical Methods

Total scores on the three measures at the baseline visit were obtained for the two samples. The three instruments have divergent ranges necessitating standardization of total scores based on the respective sample means and standard deviations. Factor analysis and BA plots were completed separately for the two cohorts.

Underlying Construct of Cognition

We subjected the total scores to a factor analysis using SPSS 15.1 (SPSS Inc., Chicago, Ill., USA) separately within each cohort in order to estimate whether a one-factor structure was reasonable (i.e., whether BA plots were appropriate) to generate independent evidence of consistent relationships among the scores. Principal component analyses, extracted from the correlation matrix, estimated the eigenvalues for the single factor we sought to extract and determined the proportion of variance accounted for by the one factor solution that we specified. If the three tests generally reflect the same underlying construct [19], one factor ('cognitive impairment') should explain the majority (at least 75%) [10, pp. 371, 372] of the variance in these three scores with factor loadings of at least 0.55 [20, p. 243]. Velicer's minimum average partial approach [21–23] was used to determine whether

Table 1. Descriptive statistics for demographic variables, including unstandardized cognitive impairment scores, the pairwise differences of their standardized values (for BA plots), and factor solution

Variable	Group	Mean $\pm$ SD percent or distribution
Age	PR	72.31 $\pm$ 7.62
	NS	73.88 $\pm$ 7.65
Sex, % female	PR	50
	NS	53
Education, years	PR	14.06 $\pm$ 3.11
	NS	13.98 $\pm$ 3.25
CDR global	PR	0.5: 31.9%; 1: 58.7%; 2: 9.4%
	NS	0.5: 34.8%; 1: 52.1%; 2: 13.1%
CDR sum of boxes	PR	5.39 $\pm$ 2.41
(0–18, higher worse)	NS	5.75 $\pm$ 2.65
ADAS-Cog	PR	21.74 $\pm$ 9.71
(0–70, higher worse)	NS	29.94 $\pm$ 7.25
MMSE	PR	8.36 $\pm$ 4.33
(0–30, higher worse) ^a	NS	9.24 $\pm$ 3.70
ADAS-MMSE ^{a, b}	PR	0 $\pm$ 0.68
	NS	–0.0001 $\pm$ 0.90
ADAS-CDR ^a	PR	0 $\pm$ 0.87
	NS	–0.002 $\pm$ 1.02
CDR-MMSE ^{a, b}	PR	0 $\pm$ 0.89
	NS	0.002 $\pm$ 0.92
ADAS loading	PR	0.907
on 1 factor ^c	NS	0.825
MMSE ^a loading	PR	0.902
on 1 factor ^c	NS	0.872
CDR SB loading	PR	0.833
on 1 factor ^c	NS	0.814

The formula for a coefficient of variation (CV) is

$$\left[\frac{\text{sd}}{\text{mean}} \right] \times 100\%.$$

^a MMSE scores were rescaled for this analysis (30 – existing score) so that higher scores on all three instruments reflected worse performance. To derive the regular MMSE mean, subtract the value given from 30.

^b These scores were standardized prior to this subtraction.

^c MMSE scores were reversed (see footnote a) and then standardized.

^d 77.66% of variance explained (PR) and 70.11% of variance explained (NS) by the 1 factor solution.

a single factor explained the majority of variability in each sample. This procedure partials out the component score variables from the original variables' correlation matrix one at a time, starting with the scores from the first component. The average of the squared partial correlations below the diagonal tends to decrease as long as useful components are being partialled out. When it starts to increase, no more components should be extracted.

Agreement

Data from the baseline visit were analyzed separately by cohort using nonparametric correlation coefficients, simple x-y scatter plots comparing scores pairwise, and a BA plot [11–13]. The BA plot represents the difference between two measures on the y axis while the average of the same two measures is plotted on the x axis of a scatter plot [12]. This is contrasted in the Results section to the simple scatterplot of pairs of scores, summarized by the correlation coefficient.

The 'summary' of data in the BA plot is reflected in reference lines for the mean difference between the two measures under consideration and the values 1 standard deviation away from the mean (i.e., mean $\pm$ 1 SD on the y axis). If the mean $\pm$ 1 SD describes an 'acceptably' small range of values, then agreement between the two measures would be characterized as 'sufficient' [11, 13]. Conversely, if mean $\pm$ 1 SD yields a very (or unacceptably) wide range, then we would conclude that the two measures do not agree sufficiently in their assessment of the underlying construct for interchangeability, or perhaps even conclude that they do not agree in terms of what is being measured.

Results

The data presented here are not intended to support conclusions about any one instrument. Importantly, all three measures (ADAS, CDR sum of box scores, MMSE) described here have extensive theoretical and/or practical histories; none of the three can be legitimately claimed to be the 'truest' measure of dementia or cognitive impairment. The correlation coefficients for the three scores within each cohort are presented in figures 1–3 with the respective scatterplots they summarize. Differences in strength of association were not subjected to inference tests because our analyses sought independent replication of results in the two cohorts (irrespective of their similarities or differences).

Table 1 presents the descriptive statistics for the demographic variables and cognitive impairment (unstandardized) test scores themselves, as well as for the pairwise differences in the standardized scores (i.e., the y axes in BA plots) and other background variables.

The data in table 1 suggest that the NS cohort had a higher average ADAS score than the PR cohort did (this was not tested for significance); the coefficient of variation [CV (standard deviation/mean) $\times$ 100%] for the PR cohort was nearly double that of the NS cohort for ADAS scores, and the PR cohort CV for MMSE was 1.3 times the CV for the scores in the NS cohort (data not shown). Since the MMSE and CDR sum of boxes is similar in the two groups, the differential correlations of ADAS with the other scores may be explained by the difference in the group CV (these CV values themselves

are not shown, but can be obtained given the data in table 1).

We fit a one-factor factor analysis solution to the raw scores to determine whether it was reasonable to conclude that the same underlying construct was assessed with the three instruments. Velicer's minimum average partial procedure [21, 22] suggested that a single factor explained the majority of variability given the correlation matrices of the three scores, suggesting that variability in all instruments can be explained by the same single underlying construct.

The one-factor solution explained 77.66% of the variance in the three scores in the PR cohort and 70.11% of the variance in the NS cohort. With three scores, no fit statistics are calculable [24]. Just 22% of the variance in the PR cohort and 30% of the variance in the NS cohort were left unexplained by the single factor. This suggests that all three instruments share the common factor, which we interpret roughly as 'cognitive impairment'. Loadings of total scores on this factor are shown in table 1.

Figure 1a shows strong and positive correlation between baseline ADAS values (y axis in fig. 1a) and (reversed) baseline MMSE scores (x axis in fig. 1a) in the PR cohort (Spearman's $r = 0.757$); a similar association is reflected in the NS cohort (fig. 1b; Spearman's $r = 0.600$).

Figure 1c and d shows the BA plots of ADAS and MMSE scores. The BA plots show the value of the difference between the two variables (or methods of measurement) on the y axis and the mean of the two variables on the x axis. Since the instruments are standardized, the vertical reference lines at zero in each plot show the average level of cognitive impairment in each cohort. The horizontal reference lines show the difference between the two measures with the zero point, indicating perfect agreement, flanked by lines 1 SD of difference away from the zero point (y axis).

Figure 1c and d shows that for greater levels of cognitive impairment (as x increases) the spread of points around the $y = 0$ line increases, suggesting that MMSE and ADAS scores agree less in persons with greater than average levels of cognitive impairment. The means-difference (BA) plots for both cohorts suggest that disagreement in ADAS and MMSE scores is not symmetric, and this asymmetry is associated with the best estimate of the underlying level of cognitive impairment (i.e., the mean of two assessments, x axis).

Figure 2a and b shows the scatter for standardized values of ADAS and CDR sum of box scores in the two co-

horts at their respective baseline visits. These figures generally reflect a weaker association between ADAS and CDR [Spearman's $r = 0.610$ (PR) and 0.472 (NS)] than was observed between ADAS and MMSE scores within the same cohort. The shapes and dispersal patterns in these scatter plots suggest that for lower CDR box scores there is a tighter correspondence with ADAS. The dispersal is greater for higher levels of both scores.

The BA plots in figure 2c and d reflect the same increasing disagreement between ADAS and CDR that was observed in the comparison of ADAS and MMSE as cognitive impairment increases. ADAS and CDR agree more for average or below-average levels of impairment and their agreement is worse for persons with above-average levels of impairment (i.e., x axis values).

Scatter plots of our transformed MMSE scores and CDR sum of box scores are shown in figure 3a and b. The correlation coefficients are similar for these two cohorts, suggesting a strong and positive association between reversed MMSE and CDR sum of box scores [Spearman's $r = 0.611$ (PR) and 0.607 (NS)], stronger than that observed for the other two pair of scores. The BA plots in figure 3c and d show the same pattern as was observed for the other two pairs of assessments, namely, as impairment increases (according to the average of MMSE and CDR), the agreement between these instruments decreases.

For average and below-average levels of impairment, the MMSE and CDR sum of boxes have the best agreement of the three pairs of comparisons; the majority of observed difference values fall within one standard deviation of the difference values.

Levine's test of equality of variance [25] in the differences (y axis values) for groups defined by the average level of impairment (i.e., x axis values < 0 , x axis values ≥ 0) were consistent in both cohorts, and for all comparisons except one. After correcting p values for multiple comparisons (three per cohort), significantly greater variability (i.e., poorer agreement) was observed for higher levels (x axis values ≥ 0) of impairment relative to lower levels (x axis values < 0). The one nonsignificant comparison was for the prednisone cohort comparison of variances in the combination of CDR and MMSE, with unadjusted $p = 0.065$. Thus, the agreement between standardized scores on these three cognitive and clinical assessments suggests that they cannot be considered interchangeable when greater-than-average impairment is present.

Fig. 1. ADAS vs. MMSE at baseline visits in two cohorts. MMSE score was reversed (30 - MMSE) to yield positive correlations with the other scores. **a** Scatter plot of standardized values of ADAS (y axis) and 30 - MMSE (x axis), PR cohort (Spearman's $r = 0.757$). **b** Scatter plot of standardized values of ADAS (y axis) and 30 - MMSE (x axis), NS cohort (Spearman's $r = 0.600$). **c** BA plot for standardized values of ADAS and MMSE, PR cohort. **d** BA plot for standardized values of ADAS and MMSE, NS cohort. **c, d** Horizontal lines represent perfect agreement ($y = 0$) ± 1 SD from zero.

Fig. 2. ADAS vs. CDR at baseline visits in two cohorts. **a** Scatter plot of standardized values of ADAS (y axis) and CDR (x axis), PR cohort (Spearman's $r = 0.610$). **b** Scatter plot of standardized values of ADAS (y axis) and CDR (x axis), NS cohort (Spearman's $r = 0.472$). **c** BA plot for standardized values of ADAS and CDR, PR cohort. **d** BA plot for standardized values of ADAS and CDR, NS cohort. **a, b** Horizontal lines represent perfect agreement ($y = 0$) ± 1 SD from zero.

Fig. 3. CDR vs. MMSE¹ at baseline visits in two cohorts. **a** Scatter plot of standardized values of CDR (y axis) and 30 - MMSE (x axis), PR cohort (Spearman's $r = 0.611$). **b** Scatter plot of standardized values of CDR (y axis) and 30 - MMSE (x axis), NS cohort (Spearman's $r = 0.607$). **c** BA plot for standardized values of CDR and MMSE¹, PR cohort. **d** BA plot for standardized values of CDR and MMSE¹, NS cohort. **c, d** Horizontal lines represent perfect agreement ($y = 0$) ± 1 SD from zero. ¹ MMSE score was reversed (30 - MMSE) to yield positive correlations with the other scores.

Discussion and Conclusions

This study of three widely used assessments of cognitive functioning in AD and aging demonstrated features of disagreement between the measures, represented by the BA (means-difference) plots, not provided by association among these measures (by correlation and scatter plot). A correlation coefficient cannot give any information about the extent to which two variables 'agree' in their estimation of the quantity of the target construct [11–13]. It is important to note that there is no information (or suggestion) about whether one or another of these instruments is more correct. These results simply suggest that the level of disagreement between any two of these instruments is dependent on the level of impairment. To our knowledge, the means-difference (BA) approach we applied to these data has not been used outside of the context of physical measurement, so there is no 'acceptable' range of differences in measurements from cognitive assessments. Importantly, if there was a linear relationship between the scores on these instruments, such that regression equations could 'create' an 'ADAS' version of a CDR box score, then the extent to which the measures disagree would be less worrisome (since it would be predictable). Our analyses show that the disagreement is not predictable, since it will depend on the underlying level of cognitive functioning. The patterns of agreement suggest that they cannot be considered interchangeable when greater-than-average cognitive impairment is present.

Spearman's 'theory of indifference of the indicator' is that any tests with 'equally high' correlation with the underlying construct will give the same indication of the amount of that construct [26]. The correlation coefficients were similar for each pair of scores in the two cohorts. Although the instruments differ in their content and one instrument (CDR) asks for descriptions of performance, rather than determining right/wrong answers to cognitive tasks, these differences only represent about 22–30% of the variability in the responses (i.e., that which the single factor did not explain). These results suggest that 'cognitive impairment', the hypothesized underlying construct, could be 'measured' equivalently by total scores on these three instruments since all had roughly equivalent loadings on the underlying factor.

The means-difference plot provides a sound method for describing the measurement of an underlying construct, even if the underlying construct is unobservable (and possibly measured with error). This was supported by the replication of the one-factor solution explaining 77.7% of variance in total cognitive performance test

scores in the PR cohort and 70.1% of this variance in the NS cohort. Observing different amounts of variance in each cohort explained by the single factor might be due to measurement error in all three instruments with respect to cognitive impairment; it is also possible that the imperfection in the fit (i.e., difference from 100% variance explained) arose due to the differences in what the instruments actually measure (apart from their common factor). A final possibility is that the instruments are used and interpreted differently for different dementia and severity levels – and neither our results nor our discussion address this. Future studies of cohorts with larger neuropsychological batteries (especially with tests that are psychometrically well-characterized in terms of the entire score range) would permit evaluations of the construct(s) being measured by the same tests across the entire score range.

Two directions for future work are suggested by our results. The first is to use modern test theory (item response theory, IRT); to analyze the items in each test would lead to estimates of the information that items on each test provide at different levels of cognitive impairment. IRT would reveal which items provide different information, or different qualities of information, about the cognitive functioning of individuals, and these items should be identified and (if possible) avoided in the assessment of cognitive functioning or impairment, whether for inclusion, exclusion, or as the outcomes of treatments [e.g., 27, 28] (NIH-funded initiatives using IRT and psychometric methods at <http://www.nihtoolbox.org/default.aspx> and <http://www.nihpromis.org/default.aspx>).

The inclusion/exclusion criteria for the two studies whose data were analyzed were intended to lead to fairly restricted ranges of individuals at baseline – those who might benefit most from the interventions being tested [16, 18]. Therefore, a second direction for future work would involve a larger sample with greater representation at the test extremes; it would be important to determine if the tests measure different constructs in individuals with scores at the extremes of their ranges [e.g., 29–30]. Support for different constructs represented by the test score extremes would be contrasted with evidence of greater variability at the extremes, but a single underlying construct represented by the entire continuum of scores (as was suggested by the replication of the single factor across samples in our analyses).

Our results suggest that while there is some 'indifference' in these indicators of cognitive impairment, this is not constant over our best estimate of the level of impairment present (i.e., the x axes in the BA plots). Our confi-

dence in these results is strengthened by their replication in two independent cohorts, and contrasted with information obtainable by correlation coefficients [31]. These results are consistent with other findings that at least some of these instruments are less sensitive to individual differences at the extremes of their ranges [e.g., 32].

Acknowledgments

The data were collected by the Alzheimer's Disease Cooperative Study under NIA Grant U01 AG10483. R.E.T. was supported in part by Grant M01RR13297-05 (National Center for Clinical Research) and is currently supported by Grant K01AG027172 from the National Institute on Aging.

References

- 1 Tractenberg RE, Weiner MF, Cummings JL, Thal LJ: Independence of changes in behavior from changes in cognition and function in persons with Alzheimer's disease: a factor-analytic approach. *J Neuropsychiatry Clin Neurosci* 2005;17:51–60.
- 2 Meulen EF, Schmand B, van Campen JP, de Koning SJ, Ponds RW, Scheltens P, Verhey FR: The seven minute screen: a neurocognitive screening test highly sensitive to various types of dementia. *J Neurol Neurosurg Psychiatry* 2004;75:700–705.
- 3 Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL: A new clinical scale for the staging of dementia. *Br J Psychiatry* 1982;140:566–572.
- 4 Morris JC: The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 1993;43:2412–2414.
- 5 Mohs RC, Rosen WG, Davis KL: The Alzheimer's Disease Assessment Scale (ADAS): an instrument for assessing treatment efficacy. *Psychopharmacol Bull* 1983;18:448–450.
- 6 Folstein M, Folstein S, McHugh P: 'Mini-Mental State': a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;12:189–198.
- 7 Leber PD: Guidelines for the Clinical Evaluation of Antidementia Drugs, First Draft. Rockville, US Food and Drug Administration, 1990.
- 8 Streiner DL, Norman GR: Health Measurement Scales: a Practical Guide to Their Development and Use, ed 3. London, Oxford Medical Publications, 2003.
- 9 Nunnally JC, Bernstein IH: Psychometric Theory, ed 3. New York, McGraw-Hill, 1994, p 120.
- 10 Stevens J: Applied Multivariate Statistics for the Social Sciences, ed 3. Mahwah, Lawrence Erlbaum, 1996.
- 11 Altman DG, Bland JM: Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983;32:307–317.
- 12 Oldham PD: Measurement in Medicine: the Interpretation of Numerical Data. London, English Universities Publishers, 1968, pp 150–151, 198–202.
- 13 Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–310.
- 14 Rosen WG, Mohs RC, Davis KL: A new rating scale for Alzheimer's disease. *Am J Psychiatry* 1984;141:1356–1364.
- 15 Thal LJ: Development of the Alzheimer's Disease Cooperative Study. *Int J Geriatr Psychopharmacol* 1997;1:6–9.
- 16 Aisen PS, Davis KL, Berg JD, Schafer K, Campbell K, Thomas RG, Weiner MF, Farlow MR, Sano M, Grundman M, Thal LJ: A randomized controlled trial of prednisone in Alzheimer's disease. *Alzheimer's Disease Cooperative Study. Neurology* 2000;54:588–593.
- 17 McKhann G, Drachman D, Folstein M, et al: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology* 1984;34:939–944.
- 18 Aisen PS, Schafer KA, Grundman M, Pfeiffer E, Sano M, Davis KL, Farlow MR, Jin S, Thomas RG, Thal LJ: Alzheimer's Disease Cooperative Study: Effects of rofecoxib or naproxen vs. placebo on Alzheimer disease progression: a randomized controlled trial. *JAMA* 2003;289:2819–2826.
- 19 Tractenberg RE, Aisen PS, Weiner MF, Cummings JL, Hancock GR: Independent contributions of neural and 'higher order' deficits to symptoms in Alzheimer's disease: a latent variable modeling approach. *Alzheimers Dement* 2006;2:303–313.
- 20 Comrey AL, Lee HB: A First Course in Factor Analysis, ed 3. Hillsdale, Lawrence Erlbaum, 1992, p 243.
- 21 Velicer WF: Determining the number of components from the matrix of partial correlations. *Psychometrika* 1976;41:321–327.
- 22 O'Connor BP: SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav Res Methods Instrum Comput* 2000;32:396–402.
- 23 Larsen RJ, Marx ML: An Introduction to Mathematical Statistics and Its Applications, ed 2. Englewood Cliffs, Prentice Hall, 1986.
- 24 Loehlin JC: Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis, ed 4. Mahwah, Lawrence Erlbaum, 2004.
- 25 Sheskin DJ: Handbook of Parametric and Nonparametric Statistical Procedures, ed 3. Boca Raton, Chapman & Hall/CRC, 2004.
- 26 Spearman C: The Abilities of Man: Their Nature and Measurement. New York, Macmillan, 1927 (reprinted: New York, AMS Publishers, 1981).
- 27 Farias ST, Mungas D, Reed BR, Cahn-Weiner D, Jagust W, Baynes K, Decarli C: The measurement of everyday cognition (ECog): scale development and psychometric properties. *Neuropsychology* 2008;22:531–544.
- 28 Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, Kuller L, Hall K, van Belle G: Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *J Clin Epidemiol* 2008;61:1018–1027.
- 29 Bland JM, Altman DG: Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–160.
- 30 Morris JC, Edland S, Clark C, Galasko D, Koss E, Mohs R, van Belle G, Fillenbaum G, Heyman A: The consortium to establish a registry for Alzheimer's disease (CERAD). 4. Rates of cognitive change in the longitudinal assessment of probable Alzheimer's disease. *Neurology* 1993;43:2457–2465.
- 31 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.
- 32 Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, Kuller L, Hall K, van Belle G: Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *J Clin Epidemiol* 2008;61:1018–1027.