Am J Neurodegener Dis 2014;3(3):158-169
www.AJND.us /ISSN:2165-591X/AJND0003371

Original Article

Validation of an Alzheimer's disease assessment battery in Asian participants with mild to moderate Alzheimer's disease

Joan HQ Shen[1], Qi Shen[2], Holly Yu[3], Jin-Shei Lai[4], Jennifer L Beaumont[4], Zhenxin Zhang[5], Huali Wang[6], Seong Yoon Kim[7], Christopher Chen[8], Timothy Kwok[9], Shuu-Jiun Wang[10], Dong Young Lee[11], John Harrison[12,13], Jeffrey Cummings[14]

[1]Pfizer Inc, Shanghai, China; [2]Clinical Science, Pfizer Inc, 500 Arcola Rd, Collegeville, PA, USA; [3]Pfizer Inc, 500 Arcola Rd, Collegeville, PA, USA; [4]Medical Social Sciences, Feinberg School of Medicine at Northwestern University Chicago, Illinois, USA; [5]Department of Neurology, Peking Union Medical College Hospital, Chinese Academy of Medical Services, Beijing 100730, China; [6]DementiaCare & ResearchCenter, Clinical Research Division, Peking University Institute of Mental Health, Beijing, China; [7]Department of Psychiatry, Asan Medical Center, University of Ulsan, Medical College Seoul, Korea; [8]Department of Pharmacology, National University of Singapore, Singapore; [9]Department of Medicine & Therapeutics (Geriatric Division), Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, China; [10]Faculty of Medicine, National Yang-Ming University School of Medicine, Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan; [11]Department of Neuropsychiatry, Seoul National University College of Medicine, Seoul, Korea; [12]Metis Cognition Ltd. Kilmington Common, UK; [13]Department of Medicine, Imperial College, London, UK; [14]Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV; Cleveland, OH: Weston, FL, USA

Received October 24, 2014; Accepted November 15, 2014; Epub December 5, 2014; Published December 15, 2014

Abstract: There is a lack of validated tools for assessing Alzheimer's disease (AD) across Asia. This study evaluates the psychometric properties of the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), Disability Assessment for Dementia (DAD), and Neuropsychological Test Battery (NTB) in Asian participants. Participants with mild to moderate AD (n=251) and healthy controls (n=51) from Mainland China, Taiwan, Singapore, Hong Kong, and South Korea completed selected instruments at several time points. Test-retest reliability was better than 0.70 for all tests. AD participants performed significantly more poorly than controls on every score. Within the AD group, greater disease severity corresponded to significantly poorer performance. The AD group test performance worsened over time and there was a trend for worse performance in AD compared to healthy controls over time. The ADAS-Cog, DAD, and NTB are reliable, valid, and responsive measures in this population and could be used for clinical trials across Asian countries/regions.

Keywords: Alzheimer's disease, validation

Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder and the major cause of dementia in the elderly. AD-related medical complications are among the most common causes of death in the elderly population [1]. Approved treatments have been developed in clinical trials conducted largely in North America. According to a report from the Institute of Medicine [2], such studies were an insufficient guide to practice as they had too few patients from non-North American countries or from different ethnic groups. As AD has become a global concern, including patients from Asia in clinical trials and translational research is important; China and other Asian countries have the highest number of people with dementia [3-5]. Yet a lack of standardized assessment tools has hindered clinical trials in this region.

Cognitive and functional instruments, such as the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) [6], Disability

Assessment for Dementia (DAD) [7], and the Neuropsychological Test Battery (NTB), including elements of the Wechsler Memory Scale [8, 9], measure the severity of AD-related symptoms and are considered important for exploring and providing evidence of treatment efficacy in research trials. However, the comparability of the psychometric properties of these instruments in Asian populations across regions has not been adequately assessed. For example, although ADAS-Cog was validated in Chinese, the sample size was small (n=39) and longitudinal data were not available [10]. Within the Chinese language, there could be dramatic differences in expressions and interpretations depending on the region.

This study evaluated the psychometric properties of the ADAS-Cog, DAD, and NTB in Asian participants with mild to moderate AD, including floor and ceiling effects, test-retest reliability, intra- and inter-rater reliability, construct validity in terms of convergent and divergent validity and discriminant validity, and the sensitivity to change during the longitudinal course of this study (approximately 78 weeks or 1.5 years).

Methods

Instruments/translation

In addition to ADAS-Cog, DAD and NTB, the Neuropsychiatric Inventory (NPI) [11], Clinical Dementia Rating-Sum of Boxes (CDR-SB) [12], Mini-Mental State Examination (MMSE) [13], and Dependence scale (DS) [14] were selected as references for validation. All instruments went through a vigorous and standardized translation process that involved forward translation, backward translation, in-country clinician review, and debriefing by native language speaking subjects, such as normal subjects and/or AD caregivers. This process was to ensure that the translated versions were not only conceptually equivalent to the original instrument but also culturally relevant and understandable to the target population in the target country. Efforts were made to ensure cultural adaptations, if necessary, were consistent across all translations. For each instrument, there were 7 linguistically validated translations to evaluate in the study, including Simplified Chinese (for mainland China), Traditional Chinese (for Taiwan, Hong Kong, and Singapore), English (for Hong Kong and Singapore), and Korean (for Korea).

Subjects

This study utilized a multicenter, longitudinal, observational design in participants with mild to moderate AD and normal cognition controls from Mainland China, Taiwan, Singapore, Hong Kong, and South Korea. After informed consent was obtained, eligible individuals entered a screening period of up to 31 days and, if eligible, then entered into the study and were evaluated over the next 78 weeks.

Eligibility criteria for all participants were 1) ages 50-85 years, 2) Rosen Modified Hachinski Ischemic (RMHI) score ≤ 4; and 3) fluency in local primary language and have at least an elementary education or equivalent. Inclusion criteria for the AD group were: 1) diagnosis of probable AD according to the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria; 2) MMSE score of 13 to 26, inclusive; 3) CDR global score ≥ 0.5; and 4) Screening visit brain magnetic resonance imaging (MRI) scan consistent with the diagnosis of AD. Inclusion criteria for the healthy controls were: 1) No significant memory complaints from normal population aged 50 to 85 years; 2) MMSE score of 21 to 30, inclusive; 3) CDR global score equal to 0, with a Memory Box score equal to 0; 4) Cognitively normal, based on absence of significant impairment in cognitive functions or activities of daily living; and 5) Normal brain MRI scan findings.

Instrument scoring

The ADAS-Cog, DAD, NTB, CDR-SB, MMSE, DS, and NPI were administered at screening, baseline, week-13, -26, -52, and -78. The NTB included the following subtests: Wechsler Memory Scale Visual-Paired Associates immediate and delayed scores [15], Rey Auditory Verbal Learning Test (RAVLT; immediate and delayed) [16], Wechsler Memory Scale -Digit Span forward and backward [15], Controlled Word Association Test (COWAT) [17], and Category Fluency Test (CFT) [18]. All scores were computed according to standard scoring instructions. Z-scores were calculated for each of the nine NTB components using the baseline mean and SD for all healthy controls with baseline scores. An 'executive function' z-score was obtained by averaging the z-scores from NTB components measuring executive function (CFT, COWAT, WMS-R-Digit Span). Signs were reve-

Table 1. Participant demographic and clinical characteristics

	Alzheimer's Disease (N=251)	Healthy Controls (N=51)	*P*-value
Age, Mean (SD)	72.0 (8.14)	63.6 (7.68)	< 0.001
Female, n (%)	143 (57.0)	21 (41.2)	0.039
Primary language, n (%)			0.0343
Simplified Chinese	123 (49.0)	18 (35.3)	
Traditional Chinese	57 (22.7)	17 (33.3)	
Korean	66 (26.3)	12 (23.5)	
English	5 (2.0)	4 (7.8)	
Education Level, n (%)			0.0231
Elementary School	74 (29.5)	5 (9.8)	
Middle School	47 (18.7)	14 (27.4)	
High School	51 (20.3)	15 (29.4)	
At least some college	79 (31.5)	17 (33.3)	
Civil Status, n (%)			0.4652
Married	202 (80.5)	46 (90.2)	
Widowed	44 (17.5)	4 (7.8)	
Divorced or Separated	4 (1.6)	1 (2.0)	
Never Married	1 (0.4)	0 (0.0)	
Domestic Situation, n (%)			0.178
Living with spouse	176 (70.1)	41 (80.4)	
Living with other family	64 (25.5)	10 (19.6)	
Living alone	11 (4.4)	0 (0.0)	
BMI, Mean (SD)	22.9 (2.33)	24.6 (3.20)	0.001
Brain MRI, n (%)			< 0.001
Normal	0 (0.0)	43 (84.3)	
Abnormal, not clinically significant	146 (58.2)	8 (15.7)	
Abnormal, clinically significant	101 (40.2)	0 (0.0)	
Missing	4 (1.6)	0 (0.0)	
CSDD Total, Mean (SD), Median (IQR)			
Informant Score (n=224, 48)	2.4 (2.49), 2 (4)	0.6 (1.76), 0 (1)	< 0.001
Participant Score (n=247, 51)	2.0 (2.40), 1 (3)	0.8 (1.27), 0 (1)	0.001
Rater Score (n=236, 51)	2.0 (2.26), 1 (3)	0.6 (0.98), 0 (1)	< 0.001
RMHIS Total Score, Mean (SD), Median (IQR)	0.6 (0.69), 0 (1)	0.4 (0.53), 0 (1)	0.075

rsed, as needed, prior to summing such that higher NTB z-scores indicate better cognitive functioning. The remaining six components, which measure memory, were averaged to obtain a 'memory' z-score (WMS-R-Visual-Paired Associates, WMS-R-Verbal-Paired Associates, RAVLT, all with immediate and delayed components).

Laboratory apolipoprotein E (ApoE) genotyping

ApoE genotypes were determined by Quest Diagnostics using QIAGEN PyroMark™ ApoE Test Kit.

Statistical analysis

Test-retest reliability of all tests were evaluated by calculating the intraclass correlation coefficient (ICC) using data from the screening and baseline assessments (25 to 31 days apart). ICCs were also calculated to evaluate the inter- and intra-rater reliability using videotaped assessments. Two AD subjects from each site were videotaped for ADAS-Cog, DAD and NTB administration at baseline visit. For intra-rater reliability, the video-recording of the baseline scale administrations was reviewed by the same raters and scored again within 7-21 days

of the live assessment. For inter-rater reliability, a rater different from the rater who performed the initial assessment viewed the video-recordings and scored the assessments within 7-21 days of the live assessment.

Spearman correlation coefficients were calculated among all scores to assess convergent and divergent validity. To assess discriminant validity, we compared mean scores between AD and control groups using analysis of covariance (ANCOVA) controlling for age and education. ANCOVA was also used to compare AD participants with mild disease (MMSE 20-26) versus moderate or severe disease (MMSE < 20), to compare AD participants across regions, and to compare ApoE4 carriers versus non-carriers.

Change from baseline was calculated for all scores. The responsiveness index (i.e., effect size), defined as the mean change in the AD groups divided by the standard deviation of the change scores in the healthy control group, was calculated to evaluate the magnitude of change overtime. We also compared mean change scores between AD and control groups with adjustment for baseline scores using ANCOVA. Longitudinal data were analyzed using mixed effects linear models for repeated measures. The mixed effects models included study group, visit and group*visit as fixed effects, controlling for other baseline covariates (age, gender, region, and education).

Results

Participants

The screening phase included 333 potential participants; 31 (29 AD, 2 healthy controls) did not complete the screening process, resulting in 251 AD and 51 healthy controls. Sites from the Chinese mainland (9 sites, 115 AD participants, 18 controls) represented nearly half of the sample, followed by those from Korea (6, 66, 12), Hong Kong (3, 33, 11), Taiwan (3, 25, 6), and Singapore (3, 12, 4). Of these participants, 208 AD and 49 healthy controls completed the entire study. Across visits, compliance with test completion ranged from 94-100%. The mean age was 70.5 (8.62). Most were female (54.7%), married (82.9%) and living with a spouse (71.4%). The years since diagnosis of AD at screening was 2.4 years (standard deviation [SD]=2.25, range: 0 to 14 years). Less than 7% of participants demonstrated any significant depressive symptoms (score ≥ 6) on any of the three CSDD scales. RMHI Scores were greater than 1 in 6.6% of participants. Most of these metrics varied in a statistically significant manner ($P < 0.05$) by study group. At baseline, after adjusting for age, gender, education, and MMSE, the scores which significantly differed ($P < 0.05$) between regions were DAD (P=0.005), CDR-SB ($P < 0.001$), and Executive Function of the NTB (P=0.008). **Table 1** summarizes other demographic and clinical characteristics by study group.

The ADAS-Cog exhibited no floor or ceiling effects on either group. In 14/17 NTB subcomponent tests, at least some AD participants scored the minimum possible, although the extent varied greatly across tests (1%-70%). Only in 4 NTB tests did some AD participants score the maximum possible (2%-18%). On the other hand, healthy controls rarely scored the minimum while in 8 of 17 NTB tests, some control participants achieved the maximum possible (8%-80%). Notably, 9% of AD participants achieved the best possible DAD score at baseline, compared to 96% of controls.

Reliability

Test-retest reliabilities: These were acceptable ICC between screening and baseline, with ICC > 0.7 for 17 of 19 measures. Two measures with ICC < 0.7 were: Wechsler Memory Scale (WMS) Visual-Paired Associates Immediate (ICC=0.5) and Delayed Memory tests (ICC=0.49). Inter- and intra-rater reliability was assessed on data from 45 videos of participants. ICC estimates were ≥ 0.91 for all except WMS Visual-Paired Associates Immediate tests where ICC=0.85 and 0.86 for within and between raters, respectively.

Convergent and divergent validity: Among non-NTB tests, 14 of 21 comparisons had Spearman's rho 0.30 or greater. NPI-caregiver distress was poorly correlated with all scales (rho: 0.14 to 0.23) except NPI (rho=0.84). When compared to the NTB, ADAS-Cog, MMSE, and CDR-SB were significantly ($p < 0.001$) correlated with Executive Function, Memory Function and Total NTB scores (rho: 0.33 to 0.71); DAD and Dependence Scale were significantly ($P < 0.001$) correlated with Executive Function and Memory Function (rho: 0.36 to 0.36); DAD and

Table 2. Analysis of covariance (ANCOVA) between Alzheimer's disease participants and healthy controls at baseline, controlling for age and education

Scale/Test	Alzheimer's Disease (N=244)		Healthy Controls (N=51)		ANCOVA				
	Mean	SE	Mean	SE	$\mu_{control}$-μ_{AD}	SD_{pooled}	Effect Size*	F-statistic	*P*-value
Disability Assessment for Dementia	77.5	1.06	98.4	2.43	20.8	16.12	1.29	59.77	< 0.001
Clinical Dementia Rating -SOB	5.1	0.15	0.1	0.35	-4.9	2.31	-2.13	162.78	< 0.001
Mini-Mental State Exam	19.7	0.23	28.9	0.54	9.2	3.57	2.58	238.4	< 0.001
Dependence Scale	5.3	0.13	0.4	0.29	-5	1.95	-2.55	232.53	< 0.001
Neuropsychiatric Inventory Total Score	7.9	0.6	1	1.38	-6.9	9.17	-0.75	20.23	< 0.001
NPI Caregiver Distress	4.2	0.31	0.4	0.71	-3.8	4.75	-0.81	23.31	< 0.001
ADAS-Cog	21.1	0.53	4.5	1.2	-16.5	7.98	-2.07	153.97	< 0.001
NTB-Executive Function	-1.2	0.04	0	0.09	1.2	0.62	1.96	137.74	< 0.001
NTB-Memory	-2.4	0.06	0.4	0.13	2.8	0.88	3.19	365.83	< 0.001
Total NTB Score	-2	0.05	0.3	0.11	2.3	0.7	3.27	382.21	< 0.001

*Effect size is defined as the absolute value of Cohen's d; d=[mean difference]/[pooled SD].

Table 3. Analysis of covariance (ANCOVA) between MMSE-derived impairment groups at screening and baseline, controlling for age and education

Scale/Test	Mild Disease (N=136)		Moderate Disease (N=108)		ANCOVA				
	Mean	SE	Mean	SE	μ_{mod}-μ_{mild}	SD_{pooled}	Effect Size*	F-Statistic	*P*-value
Disability Assessment for Dementia	84.2	1.42	68.3	1.61	-15.9	16.02	-0.99	53.72	< 0.001
Clinical Dementia Rating-SOB	4	0.2	6.4	0.23	2.4	2.27	1.06	60.83	< 0.001
Mini-Mental State Exam	22.9	0.18	15.9	0.2	-6.9	2	-3.48	658.6	< 0.001
Dependence Scale	4.6	0.17	6.3	0.2	1.8	1.96	0.9	43.77	< 0.001
Neuropsychiatric Inventory	7.3	0.89	8.6	1	1.3	9.99	0.13	0.97	0.327
NPI Caregiver Distress	3.6	0.45	4.8	0.51	1.2	5.09	0.23	2.97	0.086
ADAS-Cog	16.2	0.62	27	0.7	10.9	6.94	1.57	134.04	< 0.001
NTB-Executive Function	-1	0.05	-1.6	0.05	-0.6	0.53	-1.09	65.23	< 0.001
NTB-Memory	-2	0.07	-2.9	0.08	-0.9	0.78	-1.18	76.08	< 0.001
Total NTB Score	-1.6	0.05	-2.4	0.06	-0.8	0.6	-1.36	100.12	< 0.001

*Effect size is defined as the absolute value of Cohen's d; d=[mean difference]/[pooled SD].

Dependence Scale were significantly (P < 0.01) correlated with Total NTB scores with rho=0.19 and 0.18 for DAD and Dependence Scale, respectively.

Discriminant validity: Comparisons of demographic and clinical characteristics across groups demonstrated statistically differences in age and education. Therefore, age and education were accounted for in the following series of ANCOVA analyses. As shown in **Table 2**, AD participants performed poorer (P < 0.001) than healthy controls on all comparisons with effect sizes ranging from 0.75 (NPI) to 3.27 (total NTB score). As shown in **Table 3**, participants with moderate or severe AD showed significantly (P < 0.001) poorer performance on nearly every assessment with effect sizes ranging from 0.9 (Dependence Scale) to 3.48 (MMSE) than participants with mild AD. NPI total and caregiver distress scores did not significantly differ across AD severity levels (**Table 3**).

ApoE4 was detected in 85 of 210 (40.5%) AD participants with test results and 5 of 36 (13.9%) healthy controls (P=0.002). Among AD participants, ApoE4 was detected in 44 (40.7%) and 41 (40.2%) with mild disease and moderate/severe disease, respectively. Although there was no significant difference between ApoE4 carriers versus non-carriers in Executive Function, Memory Function, total NTB score, significant differences were found between ApoE4 carriers versus non-carriers on three NTB component tests: RAVLT Delay (P=0.024, ES=0.32), COWAT (P=0.002, ES=0.44) and Category Fluency Test (P=0.022, ES=0.33).

In terms of regions, differences in assessment scores among AD patients were found for

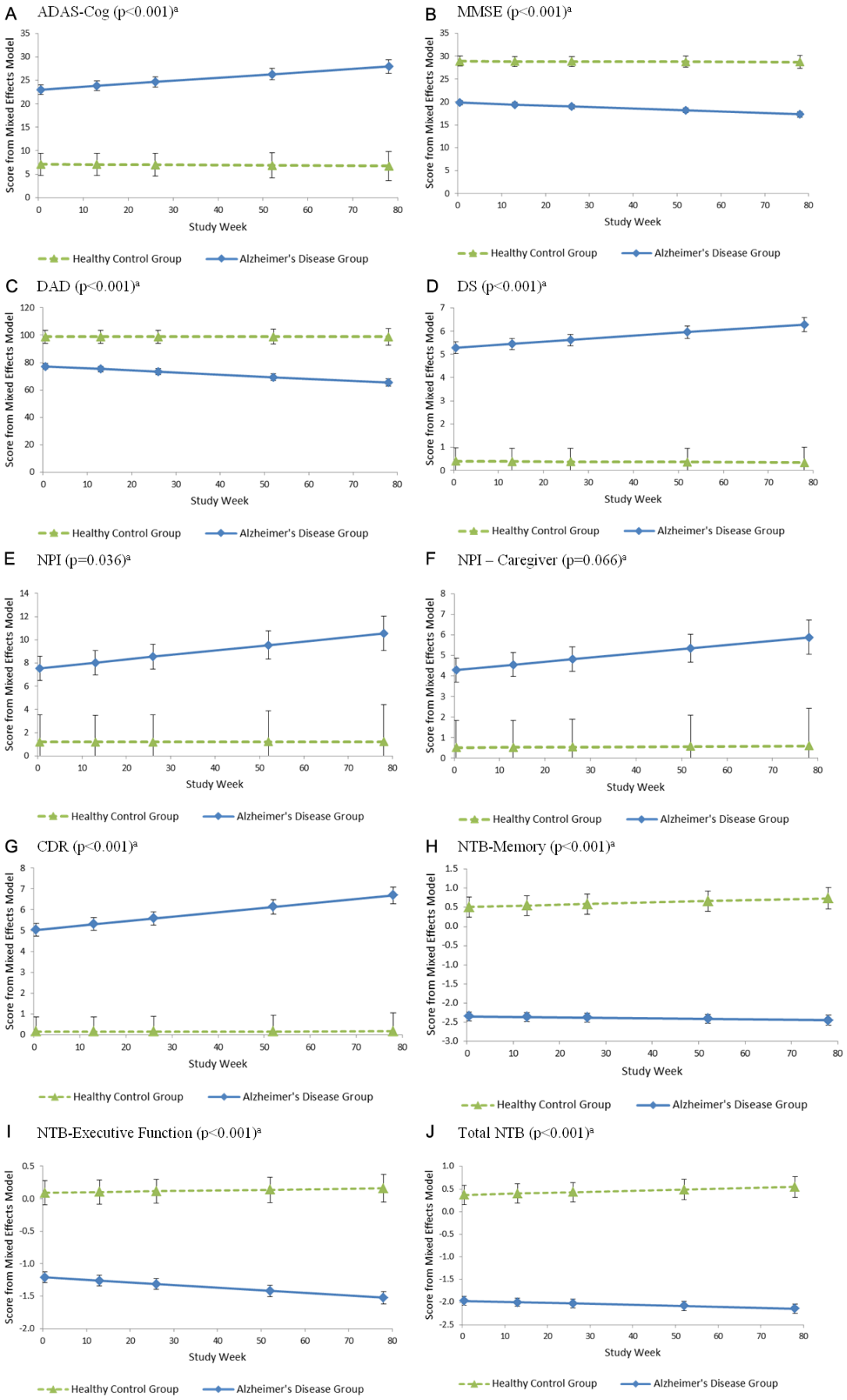
A ADAS-Cog (p<0.001)[a]
B MMSE (p<0.001)[a]
C DAD (p<0.001)[a]
D DS (p<0.001)[a]
E NPI (p=0.036)[a]
F NPI – Caregiver (p=0.066)[a]
G CDR (p<0.001)[a]
H NTB-Memory (p<0.001)[a]
I NTB-Executive Function (p<0.001)[a]
J Total NTB (p<0.001)[a]
Score from Mixed Effects Model
Study Week
Healthy Control Group
Alzheimer's Disease Group

Figure 1. Least squares adjusted means from mixed effects model including random intercept and slope terms. [a]*P*-value for the interaction between week and group, which evaluates whether the trend over time differs statistically between groups in mixed effect models, adjusting for age, gender, education, and region.

Table 4. Classification of change from baseline to week 78, based on regression based change scores incorporating age, gender, education, region, and baseline score

	Healthy Controls	AD group	*P*-value[a]	*P*-value[b]
ADAS-Cog				
Improved	1 (2%)	0	< 0.001	< 0.001
Unchanged	45 (92%)	35 (16%)		
Worsened	3 (6%)	178 (84%)		
MMSE				
Improved	1 (2%)	3 (1%)	< 0.001	< 0.001
Unchanged	45 (92%)	25 (11%)		
Worsened	3 (6%)	191 (87%)		
DAD				
Improved	0	20 (9%)	< 0.001	< 0.001
Unchanged	46 (94%)	12 (5%)		
Worsened	3 (6%)	187 (85%)		
DS				
Improved	2 (4%)	11 (5%)	< 0.001	< 0.001
Unchanged	45 (92%)	22 (10%)		
Worsened	2 (4%)	186 (85%)		
NPI				
Improved	1 (2%)	55 (25%)	< 0.001	< 0.001
Unchanged	46 (94%)	60 (27%)		
Worsened	2 (4%)	104 (47%)		
NPI Caregiver				
Improved	0	44 (20%)	< 0.001	0.005
Unchanged	44 (90%)	82 (37%)		
Worsened	5 (10%)	93 (42%)		
CDR-SB				
Improved	0	22 (10%)	< 0.001	< 0.001
Unchanged	47 (96%)	16 (7%)		
Worsened	2 (4%)	181 (83%)		
NTB-Memory				
Improved	3 (6%)	0	< 0.001	< 0.001
Unchanged	42 (86%)	73 (34%)		
Worsened	4 (8%)	143 (66%)		
NTB-Executive Function				
Improved	2 (4%)	0	< 0.001	0.016
Unchanged	44 (90%)	90 (42%)		
Worsened	3 (6%)	126 (58%)		
Total NTB				
Improved	4 (8%)	1 (< 1%)	< 0.001	0.003
Unchanged	42 (86%)	60 (28%)		
Worsened	3 (6%)	155 (72%)		

[a]Chi-squared tests comparing the change from baseline to 78 weeks between AD participants and healthy controls; [b]Difference between groups over time using GEE.

MMSE (P < 0.001), ADAS-Cog (P < 0.001), RAVLT Immediate (P=0.010), Digit Span (P < 0.001) and Executive Function (P < 0.001). Sample sizes of each region were relatively small except for mainland China (n=110, 32, 66, 11, and 25 for China, Hong Kong, Korea, Taiwan and Singapore, respectively), and this result should be interpreted with caution. The following sensitivity to change analyses were conducted adjusted for region.

Sensitivity to change

Estimated means from the longitudinal mixed effects models are plotted in **Figure 1A-J**. The primary term of interest from these models is the interaction between week and group, which evaluates whether the trend over time differs between groups. A significant interaction between group and time was observed for all measures except the NPI Caregiver scores where P= 0.066. The trend over time differed between Mainland China and Korea --- the only countries with sufficient sample size for subgroup analyses --- only in Memory Function (P=0.040) with a greater degree of change observed for China.

As shown on **Table 4**, results of the multiple-regression-based change score analysis, which incorporates age, gender, education, region, and baseline score, indicated a substantial portion of AD participants worsened to a statistically significant degree, relative to the healthy controls (all P < 0.001).

In a simplified analysis that did not adjust for demographic dif-

Table 5. Mean change from baseline to Week 78 within healthy control group and Alzheimer's disease group, adjusted for baseline score, tested against a null hypothesis value of 0

Group/Assessment	Mean change from baseline	SE	Responsiveness index[a]	*P*-value[b]	*P*-value[c]
ADAS-Cog					
Control, Normal Cognition	2.992	1.394	1.27	0.033	0.588
Mild-Moderate Alzheimer's	3.866	0.569		< 0.001	
MMSE					
Control, Normal Cognition	-1.132	0.678	-1.88	0.096	0.168
Mild-Moderate Alzheimer's	-2.217	0.263		< 0.001	
DAD					
Control, Normal Cognition	1.216	2.339	-22.31	0.604	< 0.001
Mild-Moderate Alzheimer's	-11.884	1.021		< 0.001	
DS					
Control, Normal Cognition	-0.856	0.305	2.94	0.005	< 0.001
Mild-Moderate Alzheimer's	1.123	0.115		< 0.001	
NPI					
Control, Normal Cognition	-1.781	1.452	2.3	0.221	0.001
Mild-Moderate Alzheimer's	3.513	0.668		< 0.001	
NPI caregiver					
Control, Normal Cognition	-0.448	0.77	1.8	0.561	0.008
Mild-Moderate Alzheimer's	1.84	0.354		< 0.001	
CDR-SB					
Control, Normal Cognition	0.036	0.378	16.5	0.924	< 0.001
Mild-Moderate Alzheimer's	1.647	0.15		< 0.001	
NTB-Memory					
Control, Normal Cognition	0.592	0.103	-0.14	< 0.001	< 0.001
Mild-Moderate Alzheimer's	-0.126	0.037		< 0.001	
NTB-Executive Function					
Control, Normal Cognition	0.198	0.077	-0.58	0.011	< 0.001
Mild-Moderate Alzheimer's	-0.311	0.032		< 0.001	
Total NTB					
Control, Normal Cognition	0.332	0.086	-0.42	< 0.001	< 0.001
Mild-Moderate Alzheimer's	-0.157	0.031		< 0.001	

[a]The responsiveness index is a measure of effect size calculated as the mean change in the Alzheimer's disease groups divided by the standard deviation of the change scores in the healthy control group; [b]*P* value of the mean change within group different from baseline; [c]*P* value of the difference between groups.

ferences, the mean change from baseline to Week 78, adjusted for baseline score, was evaluated within each group against the null hypothesis that the change is equal to zero (**Table 5**). Patients with AD significantly ($P < 0.001$) worsened on all scores while control group significantly ($P < 0.05$) worsened on ADAS-Cog, DS, Memory Function, Executive Function and overall NTB scores. Responsiveness index ranged from 0.14 (Memory Function) to 22.31 (DAD). Significant change score differences between AD and control groups were found on all but two (ADAS-Cog and MMSE) scores as shown on **Table 5**.

Discussions

This is the first large scale study which included multiple Asian countries/regions of psychometric properties of major AD instruments. Using data from both patients with AD and healthy controls, we verified psychometric properties of commonly used assessment tools, including acceptable test-retest reliability, inter and intra-rater reliability, validity, and responsiveness over a period of 78 weeks. Within the AD group, test-retest reliability was better than 0.70 for all tests. DAD, ADAS-Cog, DS, CDR-SB and MMSE scores correlated well with NTB component, summary and total scores, achieving significance in nearly every comparison. After adjustment for age and education differences, AD participants performed more poorly than controls on every assessment at all visits with large effect sizes. Effect sizes for NPI total and caregiver distress scores were the lowest among the assessments. Within the AD group, greater disease severity corresponded to significantly poorer performance on nearly every assessment. Only NPI total and caregiver distress scores did not significantly differ across AD participants with low versus high disease severity.

Some differences emerged between the performance of the instruments in this Asia-only cohort versus previous global studies. The

mean change in the ADAS-Cog (3.9) was somewhat less in the AD group compared to past studies (usual range 5-8). For example, recent studies of solanezumab and semagacestat, had ADAS-Cog declines of 4.5 points in Expedition 1 and 6.6 points in Expedition 2 (solanezumab) [21] and 7.8 points in the semagecestat studies [22]. Similarly, changes in DAD, NTB Total, and MMSE were smaller than observed in these prior studies. Several factors could contribute to this difference, including better overall support and adherence to treatments. A smaller change score - if reproducible - would affect power calculations and sample sizes required to show a drug-placebo difference in Asian populations.

ApoE4 was detected in significantly more AD participants than healthy controls with similar rates across regions. This finding confirms the documented literature in which ApoE4 was considered a risk factor for developing AD [23]. The rate of e4 was less than in many previously reported AD studies [24], yet similar to reports that Asian patients with AD had a lower prevalence of ApoE4 compared to US [25] and northern Europeans [26]. Regardless of its significant role in predicting AD, ApoE4 had little association with psychometric assessments, except the NTB executive function score. Our finding matched the literature [27], in which inclusion and exclusion of ApoE4 did not influence the predictive accuracy of AD progression (81% versus 80% for inclusion and exclusion, respectively).

The AD group demonstrated substantial worsening of most scores with large effect sizes represented by the responsiveness index. There was large variability in the responsiveness index for the psychometric tests, with effect sizes ranging in magnitude from 0.14 for the Memory Function (or 1.3 for non-NTB/ADAS-Cog) to 22.3 for the DAD. It should be noted that the variability in the healthy control group change scores, which constitutes the denominator of the responsiveness index, was small for most tests (ranging from 0.02 for DS and CDRS-SB to 0.35 for ADAS-Cog) and this may contribute to the larger responsiveness index values. Compared to healthy controls and adjusted for demographics and baseline score, the trend over time was significantly worse for the AD group for all measures except the NPI Caregiver scores which was expected. Yet significant NPI caregiver score differences between the AD and healthy control groups were found at all time-points.

Prior to this study, these instruments have been the subject of minimal psychometric research in Asia, although they have been shown to be valid, reliable, and responsive to change in the nations and regions in which they were developed. Specifically, prior studies have shown that the ADAS-Cog is sensitive to age-related decline in patients with mild to moderate AD [28]. The DAD has been used as an endpoint to assess functional outcomes of AD patients after treatment [7]. The individual measures of the NTB have been shown to be reliable, valid for use in AD, and sensitive to cognitive decline [29]. The NTB also evaluates delayed recall and executive function, cognitive domains that are not adequately assessed with the ADAS-Cog [30]. The NPI [11, 31], CDR [32], and MMSE [13, 33] were all developed specifically to assess patients with dementia or other cognitive impairment. The results of this study suggest that, in Asia, these instruments are also reliable, valid in differentiating cognitively impaired from cognitively healthy subjects, and sensitive in documenting longitudinal change in an AD patient group.

According to the Global Burden of Disease estimates for the 2003 World Health Report [34], dementia contributed 11.2% of years lived with disability in people aged 60 years and older. Using a Delphi consensus approach conducted by Alzheimer's Disease International, Ferri and colleagues [5] reported that although the expert consensus was for a higher prevalence of dementia in developed region, it is China and its developing western-Pacific neighbors that have the highest number of people with dementia (6 million), followed by western Europe with 4.9 million, and North America with 3.4 million. They also predicted that by 2040 China and its western-Pacific neighbors will have three times more people living with dementia than Western Europe. Zhang and colleagues reported the prevalence for persons 65 years or older was 4.8% for AD and 6.8% for dementia in China, after post-hoc correction for negative screening errors [35]. Chan and colleagues reported that the number of dementia patients were 9.19 million (5.92-12.48) in 2010 and the number of

people with AD was about 5.69 million (3.85-7.53) at the same period [36]. Catindig and colleagues [37] claimed that the dementia subtype pattern appeared to have changed over time with AD becoming more prevalent in East Asia countries since 1990. All highlighted the importance of including Asian countries in global clinical trials. The impact of dementia in Asia on health, society and the economy requires more attention. More studies using standardized cross-culturally sensitive cognitive instruments and ascertainment of functional and social declines are needed to better understand the burden and cause of early dementia [37]. The results of this study fill this important gap for Asia. Not only did it provide the validated instruments for future AD research in this region, but also provided invaluable information on how to conduct AD clinical trials in this region.

Some limitations are noted. The sample was limited to five Asian countries and therefore results cannot be generalized to other Asian countries. Additionally, although the overall sample size in this study was sufficient for psychometric validation analyses, they were not the same across all regions and only China and Korea had sample sizes greater than 50. These unequal and small sample sizes within regions limited the potential of examining the impact of cultural differences across all regions using advanced item response theory [38, 39]. To our knowledge, this is one of the first validation studies using sufficient numbers of patients across Asian regions to investigate more diverse Asian populations, allowing results to be more broadly generalizable. Future studies that recruit more diverse samples across more regions can further enhance the generalizability of the results.

In conclusion, the psychometric properties of the ADAS-Cog, DAD, and NTB were verified using data from patients with mild to moderate AD recruited from Asia. These instruments can be used for future clinical trials in the participating countries/region. Additionally, a significant amount of information was obtained, including the rates of ApoE4 carrier status in Asian AD patients which warrants further investigation. The trial was complicated and challenging, but further demonstrated the potentials and capacities of the participating sites cross countries/regions in collaboratively conducting global AD trials.

Disclosure of conflict of interest

This study was sponsored by Pfizer Inc. and Janssen Alzheimer Immunotherapy Research and Development, LLC. Jin-Shei Lai and Jennifer L. Beaumont are employees of Northwestern University, who were paid consultants to Pfizer in connection with the development of this manuscript. Joan Shen was a full-time employee of Pfizer when the study was conducted.

Address correspondence to: Dr. Joan HQ Shen, Pfizer Inc, Shanghai, China. E-mail: joanshen1@gmail.com; Dr. Qi Shen, Clinical Science, Pfizer Inc, 500 Arcola Rd, Collegeville, PA, USA. E-mail: Qi.Shen@pfizer.com; Dr. Jeffrey Cummings, Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV; Cleveland, OH: Weston, FL, USA. E-mail: cumminj@ccf.org

References

[1] DeKosky ST, Orgogozo JM. Alzheimer disease: Diagnosis, costs, and dimensions of treatment. Alzheimer Dis Assoc Disord 2001; 15 Suppl 1: S3-S7.

[2] Institute of Medicine. Initial national priorities for comparative effectiveness research. Washington, D.C.: Institute of Medicine of the National Academies; 2009.

[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: 566-72.

[4] Graham L. AAFP and ACP release guideline on dementia treatment. Am Fam Physician 2008; 77: 1173-5.

[5] Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, Hall K, Hasegawa K, Hendrie H, Huang Y, Jorm A, Mathers C, Menezes PR, Rimmer E, Scazufca M; Alzheimer's Disease International. Global prevalence of dementia: a Delphi consensus study. Lancet 2005; 366: 2112-7.

[6] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984; 141: 1356-64.

[7] Gauthier S, Gélinas I, Gauthier L. Functional disability in Alzheimer's disease. Int Psychogeriatr 1997; 9: 163-5.

[8] Wechsler D; Psychological Corporation, Pearson Education I. WAIS-IV Wechsler adult intelligence scale. 4th ed. San Antonio, Texas: Psychological Corp.; 2008.

[9] Bowden SC, Weiss LG, Holdnack JA, Lloyd D. Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale--III. Psychol Assess 2006; 18: 334-9.
[10] Chu LW, Chiu KC, Hui SL, Yu GK, Tsui WJ, Lee PW. The reliability and validity of the Alzheimer's Disease Assessment Scale Cognitive Subscale (ADAS-Cog) among the elderly Chinese in Hong Kong. Ann Acad Med Singapore 2000; 29: 474-85.
[11] Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The Neuropsychiatric Inventory: Comprehensive assessment of psychopathology in dementia. Neurology 1994; 44: 2308-14.
[12] Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology 1993; 43: 2412-4.
[13] Folstein MF, Folstein SE, McHugh PR. Mini-Mental State Examination: User's Guide. Odessa, FL: Psychological Assessment Resources, Inc.; 2000.
[14] Brickman AM, Riba A, Bell K, Marder K, Albert M, Brandt J, Stern Y. Longitudinal assessment of patient dependence in Alzheimer disease. Arch Neurol 2002; 59: 1304-8.
[15] Wechsler DS. WMS-III Wechsler memory scale - third edition. San Antonio, Texas: Psychological Corp.; 1997.
[16] King JH, Gfeller JD, Davis HP. Detecting simulated memory impairment with the Rey Auditory Verbal Learning Test: Implications of base rates and study generalizability. J Clin Exp Neuropsychol 1998; 20: 603-12.
[17] Patterson J. Controlled Oral Word Association Test. In: Kreutzer J, DeLuca J, Caplan B, eds. Encyclopedia of Clinical Neuropsychology. New York: Springer; 2011. pp. 703-6.
[18] Acevedo A, Loewenstein DA, Barker WW, Harwood DG, Luis C, Bravo M, Hurwitz DA, Aguero H, Greenfield L, Duara R. Category Fluency Test: Normative data for English- and Spanish-speaking elderly. J Int Neuropsychol Soc 2000; 6: 760-9.
[19] Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: a comparison of four models. J Int Neuropsychol Soc 1999; 5: 357-69.
[20] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13-22.
[21] Salloway S, Sperling R, Fox NC, Blennow K, Klunk W, Raskind M, Sabbagh M, Honig LS, Porsteinsson AP, Ferris S, Reichert M, Ketter N, Nejadnik B, Guenzler V, Miloslavsky M, Wang D, Lu Y, Lull J, Tudor IC, Liu E, Grundman M, Yuen E, Black R, Brashear HR; Bapineuzumab 301 and 302 Clinical Trial Investigators. Two Phase 3 Trials of Bapineuzumab in Mild-to-Moderate Alzheimer's Disease. N Engl J Med 2014; 370: 322-33.
[22] Doody RS, Thomas RG, Farlow M, Iwatsubo T, Vellas B, Joffe S, Kieburtz K, Raman R, Sun X, Aisen PS, Siemers E, Liu-Seifert H, Mohs R; Alzheimer's Disease Cooperative Study Steering Committee; Solanezumab Study Group. Phase 3 Trials of Solanezumab for Mild-to-Moderate Alzheimer's Disease. N Engl J Med 2014; 370: 311-21.
[23] Albert MS. Cognitive and Neurobiologic Markers of Early Alzheimer Disease. Proc Natl Acad Sci U S A 1996; 93: 13547-51.
[24] Beydoun MA, Beydoun HA, Kaufman JS, An Y, Resnick SM, O'Brien R, Ferrucci L, Zonderman AB. Apolipoprotein E ε4 Allele Interacts with Sex and Cognitive Status to Influence All-Cause and Cause-Specific Mortality in U.S. Older Adults. J Am Geriatr Soc 2013; 61: 525-34.
[25] Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: A meta-analysis. JAMA 1997; 278: 1349-56.
[26] Ward A, Crean S, Mercaldi CJ, Collins JM, Boyd D, Cook MN, Arrighi HM. Prevalence of Apolipoprotein E4 Genotype and Homozygotes (APOE e4/4) among Patients Diagnosed with Alzheimer's Disease: A Systematic Review and Meta-Analysis. Neuroepidemiology 2012; 38: 1-17.
[27] Fleisher AS, Sowell BB, Taylor C, Gamst AC, Petersen RC, Thal LJ; Alzheimer's Disease Cooperative Study. Clinical predictors of progression to Alzheimer disease in amnestic mild cognitive impairment. Neurology 2007; 68: 1588-95.
[28] Rockwood K, Dai D, Mitnitski A. Patterns of decline and evidence of subgroups in patients with Alzheimer's disease taking galantamine for up to 48 months. Int J Geriatr Psychiatry 2008; 23: 207-14.
[29] Harrison J, Minassian S, Jenkins L, Black RS, Koller M, Grundman M. Utility of a novel composite Neuropsychological Test Battery (NTB) for detecting cognitive change in early Alzheimer's disease patients. Alzheimers Dement 2005; 1 Suppl 1: S58.
[30] Gilman S, Koller M, Black RS, Jenkins L, Griffith SG, Fox NC, Eisner L, Kirby L, Rovira MB, Forette F, Orgogozo JM; AN1792(QS-21)-201 Study Team. Clinical effects of Aβ immunization (AN1792) in patients with AD in an interrupted trial. Neurology 2005; 64: 1553-62.
[31] Cummings JL. Changes in neuropsychiatric symptoms as outcome measures in clinical trials with cholinergic therapies for Alzheimer disease. Alzheimer Dis Assoc Disord 1997; 11 suppl 4: S1-S9. Review.

[32] Visser PJ, Verhey FR, Boada M, Bullock R, De Deyn PP, Frisoni GB, Frolich L, Hampel H, Jolles J, Jones R, Minthon L, Nobili F, Olde Rikkert M, Ousset PJ, Rigaud AS, Scheltens P, Soininen H, Spiru L, Touchon J, Tsolaki M, Vellas B, Wahlund LO, Wilcock G, Winblad B. Development of Screening Guidelines and Clinical Criteria for Predementia Alzheimer's Disease. Neuroepidemiology 2008; 30: 254-65.

[33] 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: 189-98.

[34] World Health Organization. The World Health Report 2003: Shaping the Future. Geneva: World Health Organization; 2003.

[35] Zhang ZX, Zahner GE, Román GC, Liu J, Hong Z, Qu QM, Liu XH, Zhang XJ, Zhou B, Wu CB, Tang MN, Hong X, Li H. Dementia subtypes in china: Prevalence in Beijing, Xian, Shanghai, and Chengdu. Arch Neurol 2005; 62: 447-53.

[36] Chan KY, Wang W, Wu JJ, Liu L, Theodoratou E, Car J, Middleton L, Russ TC, Deary IJ, Campbell H, Wang W, Rudan I; Global Health Epidemiology Reference Group (GHERG). Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990-2010: a systematic review and analysis. Lancet 2013; 381: 2016-23.

[37] Catindig JAS, Venketasubramanian N, Ikram MK, Chen C. Epidemiology of dementia in Asia: Insights on prevalence, trends and novel risk factors. J Neurol Sci 2012; 321: 11-6.

[38] Wright BD, Mok M. Rasch models overview. J Appl Meas 2000; 1: 83-106.

[39] Thissen D, Nelson L, Rosa K, McLeod LD. Item response theory for items scored in more than two categories. In: Thissen D, Wainer H, eds. Test scoring. Mahwah, N.J: L. Erlbaum Associates; 2001; pp. 141-86.