Alzheimer's & Dementia: Translational Research & Clinical Interventions 4 (2018) 314-323

Featured Article

TOMMORROW neuropsychological battery: German language validation and normative study

Heather R. Romero[a,b], Andreas U. Monsch[c], Kathleen M. Hayden[a,b,d], Brenda L. Plassman[a,b], Alexandra S. Atkins[e], Richard S. E. Keefe[b,e], Shyama Brewster[f], Carl Chiang[f], Janet O'Neil[g], Grant Runyan[g], Mark J. Atkinson[h,i], Stephen Crawford[h], Kumar Budur[g], Daniel K. Burns[f], Kathleen A. Welsh-Bohmer[a,b,*], for the TOMMORROW Study Investigators

[a]*Joseph & Kathleen Bryan Alzheimer's Disease Research Center (Bryan ADRC), Duke University, Durham, NC, USA*
[b]*Department of Psychiatry, Duke University, Durham, NC, USA*
[c]*University Center for Medicine of Aging, Felix Platter Hospital, Basel, Switzerland*
[d]*Department of Social Sciences and Health Policy, Wake Forest School of Medicine, Winston-Salem, NC, USA*
[e]*NeuroCog Trials, Durham, NC, USA*
[f]*Zinfandel Pharmaceuticals, Inc., Chapel Hill, NC, USA*
[g]*Takeda Development Center, Americas, Inc., Deerfield, IL, USA*
[h]*Covance Inc., Princeton, NJ, USA*
[i]*Department of Family Medicine and Public Health, University of California, San Diego, CA, USA*

Abstract

Introduction: Assessment of preclinical Alzheimer's disease (AD) requires reliable and validated methods to detect subtle cognitive changes. The battery of standardized cognitive assessments that is used for diagnostic criteria for mild cognitive impairment due to AD in the TOMMORROW study have only been fully validated in English-speaking countries. We conducted a validation and normative study of the German language version of the TOMMORROW neuropsychological test battery, which tests episodic memory, language, visuospatial ability, executive function, and attention.

Methods: German-speaking cognitively healthy controls (NCs) and subjects with AD were recruited from a memory clinic at a Swiss medical center. Construct validity, test–retest, and alternate form reliability were assessed in NCs. Criterion and discriminant validities of the cognitive measures

Conflicts of interest: H.R.R., K.M.H., B.L.P., and K.A.W.-B. received funding from Takeda as part of a contract with Duke University for the work they conducted as the neuropsychology leads to the TOMMORROW program. A.U.M. has received funding from Takeda/Zinfandel to conduct the study. He currently or in the last 3 years has received honoraria and served as a consultant or advisory board member for AbbVie, AC Immune, Lilly, Roche, and Vifor Pharma. A.S.A. is an employee of NeuroCog Trials, which provided services for this study. Currently or in the past 3 years she has received funding from the National Institute of Mental Health and the National Institute on Aging. R.S.E.K. currently or in the past 3 years has received investigator-initiated research funding support from the Department of Veteran's Affairs, National Institute of Mental Health, and the Singapore National Medical Research Council. He, currently or in the past 3 years, has received honoraria, served as a consultant, speaker, or advisory board member for AbbVie, Acadia Pharmaceuticals, Aeglea BioTherapeutics, Akebia Therapeutics, Akili, Alkermes, ArmaGen, Astellas Pharma, Avanir Pharmaceuticals, AviNeuro/ChemRar, Axovant Sciences, Biogen, Boehringer Ingelheim, Cerecor, CoMentis, Critical Path Institute, FORUM Pharmaceuticals, Global Medical Education (GME), GW Pharmaceuticals, Intra-Cellular Therapeutics, Janssen, Lundbeck, Lysogene, Medscape, Merck, Minerva Neurosciences, Mitsubishi, Monteris Medical, Moscow Research Institute of Psychiatry, Neuralstem, Neuronix, Novartis, New York State Office of Mental Health, Otsuka, Pfizer, Regenix Bio, Reviva Labs, Roche, Sangamo Therapeutics, Sanofi, Sunovion Pharmaceuticals, Takeda, Targacept, University of Moscow, University of Texas Southwestern Medical Center, and WebMD. R.S.E.K. receives royalties from the BACS testing battery, the MATRICS Battery (BACS Symbol Coding, and the Virtual Reality Functional Capacity Assessment Tool (VRFCAT). He is also a shareholder in NeuroCog Trials, which provided services for this study, and SenGenix. S.B., C.C., and D.K.B. are employees of Zinfandel Pharmaceuticals. J.O'N. is an employee of Takeda. G.R. and K.B. were employed by Takeda at the time of this study. M.J.A. was employed by Covance at the time of the study, with academic appointment at the University of California, San Diego, USA. S.C. is an employee of Covance.

*Corresponding author. Tel.: +1 919 668 1553; Fax: +1 919 669 0828.
E-mail address: kwe@duke.edu

https://doi.org/10.1016/j.trci.2018.06.009

were tested using logistic regression and discriminant analysis. Cross-cultural equivalency of performance of the German language tests was compared with English language tests.

Results: A total of 198 NCs and 25 subjects with AD (aged 65–88 years) were analyzed. All German language tests discriminated NCs from persons with AD. Episodic memory tests had the highest potential to discriminate with almost twice the predictive power of any other domain. Test-retest reliability of the test battery was adequate, and alternate form reliability for episodic memory tests was supported. For most tests, age was a significant predictor of group effect sizes; therefore, normative data were stratified by age. Validity and reliability results were similar to those in the published US cognitive testing literature.

Discussion: This study establishes the reliability and validity of the German language TOMMORROW test battery, which performed similarly to the English language tests. Some variations in test performance underscore the importance of regional normative values. The German language battery and normative data will improve the precision of measuring cognition and diagnosing incident mild cognitive impairment due to AD in clinical settings in German-speaking countries.

Keywords: Alzheimer's disease; Randomized clinical trial; Neuropsychology; Cross-cultural; Validation

1. Background

Measurement of cognitive change in Alzheimer's disease (AD) trials has typically relied on well-validated metrics, such as the Alzheimer's Disease Assessment Scale-Cognitive Subscale test, a composite measure that detects treatment response across different trial settings and cultures [1]. However, as more research started focusing on the preclinical stages of AD, more sensitive metrics are required for the detection of very subtle cognitive changes that occur in the very early stages of incident mild cognitive impairment due to AD (MCI-AD). An increasing number of secondary prevention studies are now underway that are designed to test therapeutic agents that may either delay or prevent the onset of early clinical symptoms of AD [2]. These studies have selected composite end points capable of detecting cognitive decline in older adults at risk of developing AD, based on neuropsychological data from longitudinal studies of clinically healthy populations [3,4].

The TOMMORROW study is among these secondary prevention studies and examines the efficacy of medicines to delay the onset of diagnosable MCI-AD [5,6]. Unlike the other secondary prevention trials, the primary outcome is a clinically definable event, MCI, rather than a cognitive composite. The trial uses a time-to-event design, with the primary end point event defined as an adjudicated clinical diagnosis of MCI-AD based on the National Institute on Aging–Alzheimer's Association criteria [7]. This end point choice has advantages in being a clinically meaningful outcome but also presents some complexities when applied across different languages and cultures. First, the MCI-AD criteria had not yet been used as an end point in clinical trials, and the criteria needed to be operationalized to allow diagnostic standardization across multiple sites and clinicians. Second, because the diagnostic criteria are to be used in a global context, cross-cultural validation and standardization of the metrics are essential to ensure that they capture the cognitive and functional features of early-stage AD across all the cultural settings.

In the TOMMORROW study, the core clinical criteria for MCI-AD were operationalized. These criteria were made available for public review before study launch [6] and are defined as a sustained decline in both function and neuropsychological performances. Functional decline is operationalized as a change in the Clinical Dementia Rating Scale (CDR) score from 0 to 0.5, and neuropsychological performance is measured by two domains of cognition (one of which must have been episodic memory) falling below baseline performance (to at least 1.3 standard deviation [SD] below the age-adjusted normative mean) or an isolated decline from baseline in episodic memory (i.e., to at least 1.5 SD below the age-adjusted normative mean). To meet the operationalized definition of MCI-AD, these criteria are expected to be met at two consecutive study visits, approximately 6 months apart.

Criteria for MCI-AD can be applied readily in English-speaking countries using standardized cognitive assessments that are normed for the population and well validated for clinical use in cognitively healthy older adults. However, the use of these tools for diagnostic purposes in other cultures requires establishing their test reliability, validity, and normative ranges appropriate to the applicable culture. The purpose of the present study was to psychometrically validate the German language adaptation of the cognitive tests for the TOMMORROW study for application among German-speaking adult populations aged 65–88 years (the target age range for the TOMMORROW study) and to develop normative data for the battery of tests.

2. Methods

We conducted a validation and normative study of the German language version of the TOMMORROW study's neuropsychological test battery. The primary objectives of

the study were to: (1) assess criterion validity; (2) examine the construct validity and cross-cultural equivalency of translated test results compared with US normative results; (3) assess the discriminant validity and receiver operator characteristics of cognitive tests when differentiating between cognitively healthy controls (NCs) and individuals with AD in the elderly population; (4) determine the test-retest and alternate forms reliability of select cognitive tests over a 1-month interval using NCs; and (5) establish demographically adjusted normative data for each of the instruments and selected alternative forms in the cognitive test battery in cognitively normal German-speaking elderly individuals.

The study was conducted in accordance with the Declaration of Helsinki. Procedures were approved by the local institutional review boards or ethics committees. All subjects provided informed, written consent.

2.1. Study design

2.1.1. The TOMMORROW neurocognitive battery

The TOMMORROW neuropsychological test battery (TOMMORROW battery) was designed for early detection of MCI-AD within the TOMMORROW study. Briefly, the TOMMORROW battery was designed to include sensitive tests of episodic memory, language, visuospatial ability, executive function, and attention (Table 1). Tests included the Brief Visuospatial Memory Test–Revised (BVMT-R); California Verbal Learning Test–II (CVLT-II) [8,9]; Trail Making Test Parts A and B [10]; semantic fluency [11,12], lexical fluency, the Multilingual Naming Test (MiNT); digit span forward and backward [13]; and the clock-drawing test (CDT) [14].

2.1.2. Translation of the TOMMORROW neuropsychological test battery and rater training

Linguistic and cultural adaption of the TOMMORROW battery was completed in accordance with the International Society for Pharmacoeconomics and Outcomes Research guidelines [15]. Formal permission for test use and translation was obtained through licensing agreements with each specific test publisher or copyright owner. Formal forward and backward translation of each measure was followed by in-country cognitive debriefing interviews, expert reviews, and harmonization activities to ensure that the German translations of tests were culturally appropriate [15]. Finally, to ensure the high quality of cognitive data, professional cognitive testing firms were engaged to provide rater training and data quality review services throughout the trial. Owing to known variability in scoring for the BVMT and CDT assessments, these instruments received centralized scoring.

Table 1
TOMMORROW neuropsychological test battery (key indicators)

Cognitive domain	Cognitive test/subtest
Memory	CVLT-II long-delay free recall correct [8,9]
	CVLT-II short-delay free recall correct
	BVMT-R delayed recall
Executive function	Trail Making B test (total seconds) [10]
	Digit span backward [13]
Language	Semantic fluency (animals) [11,12]
	Lexical fluency (total words)
	MiNT visual naming test
Attention	Trail Making A test (total seconds)
	Digit span forward
Visuospatial function	Clock-drawing test [14]
	BVMT-R copy accuracy

Abbreviations: BVMT-R, Brief Visuospatial Memory Test–Revised; CVLT, California Verbal Learning Test; MiNT, Multilingual Naming Test.

2.1.3. Selection of study participants

Participants were NC individuals and patients with AD, recruited from the University Center for Medicine of Aging, Felix Platter Hospital, Basel, Switzerland. This research site was selected because it would become a site for the TOMMORROW study, and the predominant language spoken (high German) is linguistically equivalent to German-speaking populations in Germany, thereby assuring generalizability of the normative data to a broad range of German-speaking older adults.

Participants were included in the NC group if they had a CDR global rating of 0, a Mini–Mental State Examination [16] age- and education-adjusted score $\geq$25 and were deemed cognitively healthy on neurological evaluation. Mini–Mental State Examination scores for NC participants were adjusted for low education (+1 point for those with 0–9 years of education) and older age (+1 point for >75 years). Participants with AD were required to have a global CDR rating between 0.5 and 2.0, and have a clinical diagnosis of possible or probable AD based on well-established diagnostic criteria and supported with clinical documentation from medical records. All participants were required to have a knowledgeable project partner to complete the CDR and other informant-based questionnaires. To ensure consistency in diagnostic assignment, an expert clinical review of cases was conducted by a clinical consensus committee who had designed the assessment battery and validation study (i.e., the Neuropsychology Lead Office for the TOMMORROW program at Duke University: K.M.H., B.L.P., H.R.R., and K.A.W.-B., in addition to A.S.A. from NeuroCog Trials) and were involved in the cultural validation (A.S.A.).

There were 50 NC participants enrolled in the following four age strata (65–69, 70–74, 75–79, and 80–88 years). To ensure a balanced representation of gender and education in each age group, enrollment per stratum was capped to include the following: (1) a minimum of 15 and at most 35 male or female participants per age stratum; (2) a minimum of five and at most 10 individuals with a lower educational level; and (3) a minimum of 40 and at most 45 individuals with a higher educational level. Definitions for high versus low education level were based on population

statistics for older adults in Switzerland, in consultation with local experts. Educational levels for Switzerland were designated as lower level (0–9 years) and higher level (grade 10 or greater). Finally, to obtain an adequate distribution of AD participants across the overall age range (65–88 years), 17 patients with AD aged 65 years to <78 years and eight patients with AD aged ≥78 years were recruited.

2.1.4. Schedule of assessments

All participants (NC and AD) and their project partners attended visit 1 (baseline). During this visit, participants were administered the Mini–Mental State Examination [16] for eligibility purposes, followed by a clinical evaluation, which included the TOMMORROW neuropsychological battery. The participants completed a depression screen (Geriatric Depression Screen [17]) and a meta-memory questionnaire (Mail-in Cognitive Function Screening Instrument [18]). Medical history was reviewed along with a full clinical interview for dementia with the participant and the project partner using the CDR scale. Clinical diagnoses were assigned by the site clinician based on all data collected at visit 1, with the exception that the clinician remained blinded to the TOMMORROW battery scores to avoid contamination of the validation process. The AD participants were administered standard forms of the CVLT-II and BVMT-R tests. The NC participants were randomized to receive either schedule 1 of the BVMT-R and CVLT-II (standard forms administered at both visits) or schedule 2 (standard forms administered at visit 1, alternate forms at visit 2).

Only the NC participants returned for visit 2 (1 month after visit 1) to assess test-retest and alternative form reliability of the TOMMORROW battery. The questionnaires assessing depression and meta-memory were readministered at visit 2 to ensure that there were no changes in cognitive state over the 1-month interval that would invalidate the individual as a normal control.

2.2. Analytic strategy

2.2.1. Overview

We first evaluated the reliability and validity of the culturally and linguistically adapted German translations of the cognitive tests in the following ways: (1) criterion validity was evaluated using analysis of covariance to compare the differences in scores on the cognitive tests at visit 1 between AD and NC participants when covarying by age strata, educational level, and gender; (2) construct validity was examined using Pearson's correlations (if normally distributed) or Spearman's rho correlations (if either distribution was nonnormal) among selected cognitive tests; (3) discriminant validity was evaluated using test classification statistics (e.g., sensitivity and specificity) for the core cognitive domains of AD; and (4) test-retest reliability was assessed using Pearson's correlation coefficients. Cross-cultural equivalence was evaluated by comparing the SDs and the 7th and 10th percentile levels of standardized scores of NC German language participants with published information in commercial manuals for the English versions of the tests (Table 1). The second set of analyses involved the generation of age-standardized German language norms.

2.2.2. Validation analysis

For discriminant validity analyses, composite scores were calculated for four cognitive domains, including episodic memory (BVMT-R delayed, CVLT-II long delay, and CVLT-II short delay), executive function (Trails B and digit span backward), language (semantic and lexical fluency), and attention (Trails A and digit span forward). Z-scores for tests within each domain were calculated and then averaged to produce the domain composite scores. The four domain scores were then averaged to produce the total neuropsychological composite score. A forced entry logistic regression was used to evaluate the performance of the composite score for discriminating between NC and AD groups.

A forward stepwise logistic regression analysis was used to identify the subset of cognitive tests that best described differences between NC and AD cases. Classification statistics (i.e., sensitivity, specificity, positive predictive value, negative predictive value, and correct classification) were generated for each set of the models at a prespecified cut point typically applied to define dementia (−2 SD). Cognitive tests were compared by domain-specific and composite models (with and without covariate adjustment) to determine the extent to which performance accurately classified diagnostic groups.

Test-retest reliability was assessed by examining the correlations between paired responses to standard forms of the cognitive tests at visit 1 and visit 2 data from NC participants. Alternate forms reliability was conducted on the raw scores from standard and alternate forms of memory tests that were administered to the same individuals (in schedule 2) at visit 1 and visit 2, respectively.

2.2.3. Sample size justification

The sample size of 50 NC participants per age strata was based on published guidelines for stable estimates in normative samples [19,20]. For validity analysis, an effect size was estimated based on previously reported mean baseline differences between NC and premorbid AD "converters" (within 3 years), equivalent to an effect size of $d > 1$ on both the Trails B and the CVLT-II tests [21]. Although conservative for this study in which we examine separation of AD dementia from NC, we expected that some tests in the battery would be less sensitive than others. Power analysis indicated that 21 patients with AD were required to detect a group difference equivalent to an effect size of $d = 0.9$ between AD and NC individuals, assuming use of a two-tailed test with $P < .05$ and powered at $\beta = 0.8$. Given the potentially smaller effect size estimates for less sensitive tests in the battery, the enrollment target was increased to 25 subjects with AD.

2.2.4. Creation of age-standardized norming tables

A series of multiple linear regression analyses were performed to evaluate the relative contribution of age strata, years of education, and gender in predicting the raw cognitive test scores of NC individuals. Age-standardized normative data tables were generated using the NC data set at visit 1, for each of the four age strata. Norming tables for each cognitive test were based on one of the following norming methods: (1) Z-scores; (2) percentiles; and (3) cutoff points. Test scores for each of the cognitive tests were normed on a linear Z-score metric with a mean of 0 and SD of 1. For tests with skewed distributions, either percentile associated with each raw score for each age strata were generated, or cutoff values were determined to represent the seventh percentile for each age strata.

3. Results

3.1. Sample

There were 228 participants screened; five did not meet the eligibility criteria. The final analysis sample consisted of 198 NC participants and 25 AD participants (Figure 1). Among the 198 NCs, 119 participants were randomized to receive schedule 1 of the BVMT-R and CVLT-II tests, and 79 were randomized to receive schedule 2. Following visit 1, one NC participant voluntarily withdrew from the study, which reduced the NC sample available for reliability analyses to 197 participants. Sample characteristics are presented in Table 2.

3.2. Cognitive test performance of NC participants

Cognitive test performance for NC participants for visit 1 and 2 are listed in Supplementary Table 1. Measurements of central tendency are provided in Supplementary Table 2.

3.3. Criterion (known groups) validity

Table 3 shows the difference in test performance for the AD and NC participants. The largest effect sizes were associated with episodic memory: CVLT-II long-delay free recall correct, CVLT-II short-delay free recall correct, and BVMT-R delayed recall. Age was a significant predictor of group effect sizes for most tests. In contrast, significant education and gender effects were found less frequently.

3.4. Construct validity

Concurrent and divergent validity were demonstrated. Correlations within highly related domains were high

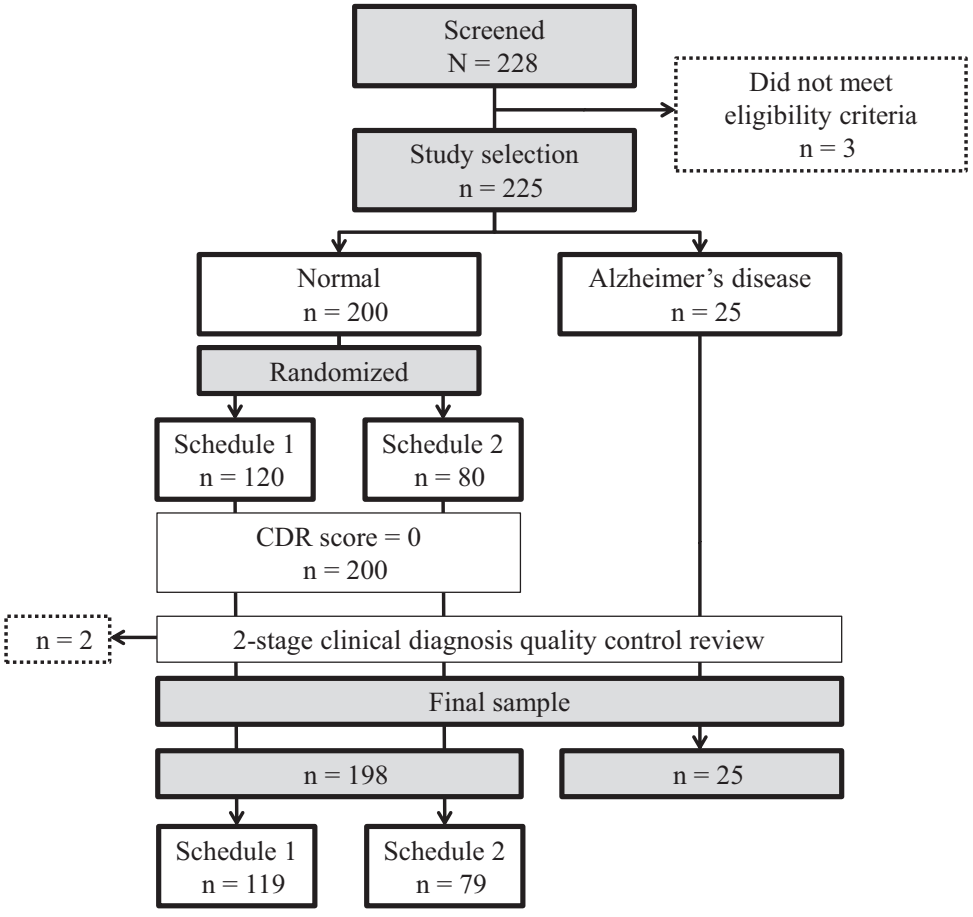

Fig. 1. Sample characteristics and study design. Abbreviation: CDR, Clinical Dementia Rating.

Table 2
Sociodemographic characteristics of study samples

Variable	NC participants, visit 1 (n = 198)	Diagnosed AD participants, visit 1 (n = 25)
Gender, n (%)		
Male	100 (50.5)	6 (24.0)
Female	98 (49.5)	19 (76.0)
Age (years)		
Mean (SD)	74.4 (5.94)	75.6 (4.93)
Median	75.0	76.0
Age category, n (%)		
65–69 years	50 (25.3)	3 (12.0)
70–74 years	48 (24.2)	4 (16.0)
75–79 years	50 (25.3)	13 (52.0)
80–88 years	50 (25.3)	5 (20.0)
Education (years)		
Mean (SD)	13.4 (2.63)	11.6 (2.40)
Median	13.0	11.0
Educational level, n (%)		
Lower level	38 (19.2)	13 (52.0)
Higher level	160 (80.8)	12 (48.0)
Race/ethnicity, n (%)		
White (not Hispanic)	198 (100.0)	25 (100.0)
MMSE, visit 1		
Mean (SD)	29.7 (1.19)	24.5 (3.20)
Minimum, maximum	25, 32	17, 28
CDR score, n (%), visit 1		
0 (normal)	198 (100.0)	0
0.5 (questionable dementia)	0	10 (40.0)
1 (mild dementia)	0	13 (52.0)
2 (moderate dementia)	0	2 (8.0)

Abbreviations: AD, Alzheimer's disease; CDR, Clinical Dementia Rating; MMSE, Mini–Mental State Examination; NC, cognitively healthy controls; SD, standard deviation.

NOTE. Educational levels are 0–9 years for lower level and 10 or greater for higher level for Switzerland.

(0.41–0.86), for example, between CVLT delayed recall and total learning or between digit span forward and total digit span (Supplementary Table 3). Weaker correlations (0.31–0.46) were observed across multicomponent domains and unrelated cognitive domains (e.g., BVMT copy to digit span).

3.5. Discriminant validity

The total composite score correctly classified NC and AD cases with 76% sensitivity and 97% specificity. The episodic memory composite score performed the best, with a sensitivity of 92% and specificity of 99%. There was some variability in the performance of the other three domain scores, with sensitivity ranging from 48.0% to 92.0% and specificity ranging from 93.4% to 99.0%. Examination of the pseudo R^2 of logistic models for each domain suggests that measures of the episodic memory domain had almost twice the predictive power of any other domain.

3.6. Test-retest and alternate form reliability

Test-retest reliability was examined over a 1-month interval in NC participants assigned to schedule 1 (n = 119). A mild practice effect was observed on cognitive test scores between visit 1 and visit 2. Eleven out of the 13 tests had correlations between visits 1 and 2 that were greater than the hypothesized threshold value of 0.60 for adequate test-retest reliability for cognitive tests (Supplementary Table 4). The remaining two tests fell below the hypothesized threshold: the CDT (0.54) and BVMT-R copy (0.40). Alternate form reliability coefficients were high and fairly consistent across tests, ranging between 0.55 and 0.71 over a 1-month interval among normal participants assigned to schedule 2 (n = 78).

3.7. Creation of age-standardized norming tables

3.7.1. Evaluation of raw score covariates: age, gender, and education

A series of multiple linear regression analyses were performed to evaluate the relative contribution of age strata, years of education, and gender in predicting the raw cognitive test scores of NC individuals. Age was a significant predictor of test scores on almost half of the tests administered, whereas fewer cognitive test scores were significantly influenced by education and gender. The impacts of age, education, and gender were test-specific (Table 4). The size of the age association, as measured by partial R^2 values was strongest for BVMT-R total recall, BVMT-R delayed recall, and Trails B. The education effect on cognitive test performance was weaker as indicated by only one strong effect on verbal learning (CVLT-II trials 1–5).

The need for gender correction was further explored for tests that had a significant gender effect ($P < .001$) including MiNT, CVLT-II short-delay free recall, CVLT-II long-delay free recall, and CVLT-II total learning tests. Several factors were taken into consideration, including the intended use of the tests in preclinical AD trials (e.g., is the test part of trigger criteria for diagnosis of MCI-AD?) and the ease of interpreting demographically corrected tests by site clinicians in a clinical trial (e.g., can the clinician interpret performance if points are added/subtracted for a specific gender?). The magnitude of differences by gender was examined across age groups in order to determine which tests should receive a gender correction. The decision was made not to apply a gender correction to MiNT and CVLT-II total learning.

Further modeling was done to examine the impact of a gender correction for CVLT-II short-delay free recall and long-delay free recall. The frequency of normal participants who scored −1.3 and −1.5 SD below the mean was examined, as well as the raw score differences between men and women at each of these cut points (Supplementary Table 5). In general, across the episodic memory tests, the gender difference in mean score was stronger in the youngest three age groups than in the oldest age group (80–88 years); however, the difference at −1.3 and −1.5 SD below the mean was more pronounced in the oldest group. An effort was made to choose an adjustment that would correct for the distinct effects of gender on CVLT-II long-delay and

Table 3
Differences in cognitive test performance: AD and NC groups, accounting for age, education, and gender

Cognitive domain	Cognitive test	NC participants (n = 198)	AD participants (n = 25)	ANCOVA, F-Statistic†	Cohen's d	Standardized β weights and sign					
						Age (years)		Education (years)		Female versus male	
Memory	CVLT-II long-delay free recall correct	10.3 (0.20)	1.3 (0.58)	67.32	3.21	−0.09	**	0.27	***	1.78	***
	CVLT-II short-delay free recall correct	9.6 (0.21)	1.6 (0.62)	46.18	2.65	−0.10	*	0.20	*	1.63	***
	BVMT-R delayed recall	7.7 (0.17)	2.0 (0.48)	43.49	2.44	−0.12	***	0.13	—	−0.15	—
Executive function	Trails B total seconds	112.1 (3.56)	230.7 (11.35)	40.49	2.38	3.19	***	−4.35	*	−6.72	—
Language	Semantic fluency (animals)	22.1 (0.32)	12.4 (0.92)	31.91	2.16	−0.13	*	0.30	*	0.96	—
Attention	Trails A total seconds	41.4 (1.84)	95.5 (5.32)	26.39	2.09	0.38	—	−0.80	—	−1.77	—
Language	MiNT	29.6 (0.18)	24.4 (0.52)	34.95	2.06	−0.07	*	0.06	—	−1.44	***
Visuospatial	Clock-drawing test	8.6 (0.11)	6.2 (0.32)	14.40	1.60	0.00	—	0.05	—	0.29	—
Language	Lexical fluency (total words)	36.1 (0.75)	25.6 (2.18)	8.08	0.99	−0.15	—	0.63	*	2.37	—
Executive function	Digit span backward	6.2 (0.13)	4.4 (0.37)	10.02	0.99	−0.05	*	0.06	—	−0.44	—
Visuospatial	BVMT-R copy accuracy	11.2 (0.08)	10.3 (0.23)	4.92	0.79	−0.02	—	0.04	—	0.20	—
Executive function	Digit span total	15.0 (0.23)	12.6 (0.65)	8.73	0.76	−0.10	—	0.15	—	−0.79	—
Attention	Digit span forward	8.8 (0.13)	8.2 (0.38)	4.38	0.33	−0.05	*	0.09	—	−0.35	—

Abbreviations: AD, Alzheimer's disease; ANCOVA, analysis of covariance; BVMT-R, Brief Visuospatial Memory Test–Revised; CVLT, California Verbal Learning Test; MiNT, Multilingual Naming Test; NC, cognitively healthy controls.
*$P < .05$; **$P < .01$; ***$P < .001$.
†All P values <.001 for differences between AD and NC groups.

short-delay recall tests, without over-correcting the scores of individuals in the oldest age group, resulting in an adjustment of +1.0 for men, to level the differences between males and females.

3.7.2. *Age-standardized norming tables*

Age-standardized normative data tables were generated using the NC data set at visit 1 for each of the four age strata. The method used for generation of normative data for each test is summarized in Supplementary Table 6, and normative tables for each test are provided in Supplementary Tables 7–9.

4. Discussion

This study established the reliability and validity of the German language TOMMORROW battery. Overall, the battery performed as expected in German speakers in Switzerland and similar to the tests in English in the US. However, there were some differences observed across the language versions, which underscore the importance of having regionally based, language-appropriate normative values. To this end, we provide normative data to facilitate the application of the battery in German speakers enrolled in TOMMORROW and other clinical trials in the future. Because all of the cognitive measures in the TOMMORROW neuropsychological battery are scaled for use in the clinical assessment of NC adults, they are well targeted for use in trials designed to prevent or delay early symptomatic AD, and possibly other mild cognitive disorders. The co-norming of the measures that was done in this study provides some unique advantages when drawing clinical inferences. Standardized scores (such as Z-scores, t-scores, or percentiles) for all measures are based on the same normative sample, a situation which allows for a ready assessment of relative strengths and weaknesses in cognitive performance. This information can bolster diagnostic confidence of suspected focal presentations (e.g., stroke), clarify profiles of generalized cognitive disorders due to various dementias (such as AD), and reinforce inferences of possible improvement or decline over time.

The TOMMORROW battery was designed to reliably capture incident cases of MCI-AD. The predetermined cut points on the composite score (−2 SD below the norm) yielded variability in sensitivity across the four domain scores; however, specificity was high for all domains. The strong discriminant performance of the CVLT-II delay tests and weaker contribution of tests of attention (e.g., digit span forward and Trails A) is consistent with findings in the cognitive testing literature that focuses on detection of mild cases of AD [22–25]. This variability across domains is expected in AD, in which memory impairment is a distinguishing feature throughout the disease, whereas the other domains change over the course of disease expression. We have constructed the composite taking into account the variability in disease expression, aiming to make the composite maximally useful in detecting longitudinal change and potential treatment effects in the very early stages of MCI-AD, rather than cross-sectional change. These expectations are supported by modeling work with the Women's Health Initiative Memory Study [26]. Additional work should be completed to further improve the composite weighting to allow optimal detection of treatment change over short- and long-time intervals anticipated in clinical prevention trials.

Tests translated and culturally adapted for the German language in Switzerland were compared with the US-developed

Table 4
Strength of age, gender, and education as predictors of NC raw cognitive test scores (n = 198)

	Age (years)			Education (years)			Gender (female vs. male)			Total	
Cognitive tests	β	Partial R^2	*P* value	β	Partial R^2	*P* value	β	Partial R^2	*P* value	R^2	*P* value
BVMT-R total recall trials 1–3 scores	−0.310	0.0912	***	0.380	0.0251	*	0.100	0.0001	—	0.1207	***
BVMT-R delayed recall scores	−0.130	0.0941	***	0.130	0.0165	*	−0.240	0.0020	—	0.1160	***
BVMT-R recognition hits scores	−0.010	0.0072	—	0.000	0.0001	—	−0.050	0.0015	—	0.0084	—
BVMT-R recognition false-positives scores	0.010	0.0448	**	−0.020	0.0146	—	0.050	0.0050	—	0.0687	**
BVMT-R recognition discrimination scores	−0.020	0.0269	*	0.020	0.0037	—	−0.100	0.0041	—	0.0355	—
BVMT-R copy accuracy scores	−0.020	0.0140	—	0.030	0.0056	—	0.150	0.0045	—	0.0248	—
Digit span forward scores	−0.060	0.0347	—	0.090	0.0137	—	−0.420	0.0103	—	0.0657	**
Digit span backward scores	−0.060	0.0321	**	0.030	0.0016	—	−0.480	0.0152	—	0.0481	*
Digit span total scores	−0.120	0.0445	**	0.120	0.0085	—	−0.900	0.0168	—	0.0737	**
Trails A total seconds scores	0.450	0.0343	**	−0.290	0.0025	—	−1.880	0.0037	—	0.0438	*
Trails A total error scores	0.000	0.0000	—	−0.010	0.0032	—	−0.010	0.0002	—	0.0033	—
Trails B total seconds scores	3.200	0.1278	***	4.020	0.0358	**	−5.950	0.0028	—	0.1728	***
Trails B total error scores	0.050	0.0667	***	−0.050	0.0123	—	0.020	0.0000	—	0.0822	***
MiNT scores	−0.060	0.0371	**	0.090	0.0166	*	−1.290	0.1142	***	0.2012	***
Semantic fluency (animals) scores	−0.150	0.0358	**	0.280	0.0221	*	1.050	0.0111	—	0.0690	**
Lexical fluency (total words) scores	−0.120	0.0046	—	0.600	0.0191	—	2.150	0.0087	—	0.0284	—
Lexical fluency (letter D) scores	−0.030	0.0012	—	0.370	0.0404	**	0.930	0.0092	—	0.0438	*
Lexical fluency (letter S) scores	−0.070	0.0085	—	0.170	0.0093	—	0.460	0.0025	—	0.0199	—
Lexical fluency (letter F) scores	−0.030	0.0019	—	0.070	0.0017	—	0.760	0.0076	—	0.0112	—
Clock-drawing test scores	0.000	0.0000	—	0.040	0.0064	—	0.290	0.0101	—	0.0128	—
CVLT-II trial 1–5 correct scores	−0.340	0.0457	**	0.950	0.0637	***	5.700	0.0807	***	0.1773	***
CVLT-II trial B scores	−0.040	0.0197	*	0.140	0.0357	**	0.990	0.0654	***	0.1110	***
CVLT-II short-delay free recall correct scores	−0.100	0.0343	**	0.170	0.0162	—	1.680	0.0553	***	0.1095	***
CVLT-II long-delay free recall correct scores	−0.100	0.0382	**	0.250	0.0397	**	1.850	0.0749	***	0.1467	***
CVLT-II primacy scores	0.050	0.0023	—	−0.270	0.0118	—	1.360	0.0103	—	0.0346	—
CVLT-II middle scores	−0.110	0.0082	—	0.400	0.0209	*	−0.090	0.0000	*	0.0325	—
CVLT-II recency scores	0.050	0.0024	—	−0.130	0.0034	—	−1.230	0.0099	—	0.0149	—
CVLT-II total intrusions scores	0.120	0.0115	—	−0.140	0.0029	—	−0.860	0.0040	—	0.0193	—
CVLT-II total repetitions scores	−0.030	0.0047	—	0.030	0.0005	—	0.150	0.0005	—	0.0062	—
CVLT-II total recognition hits scores	0.010	0.0005	—	0.040	0.0054	—	0.290	0.0084	—	0.0104	—
CVLT-II total recognition false-positives scores	0.050	0.0114	—	−0.200	0.0306	*	−1.100	0.0329	*	0.0663	**

Abbreviations: BVMT-R, Brief Visuospatial Memory Test–Revised; CVLT, California Verbal Learning Test; MiNT, Multilingual Naming Test; NC, cognitively normal healthy controls.
*$P < .05$; **$P < .01$; ***$P < .001$.

versions. The validity, test-retest reliability, and alternate form reliability results all showed that the psychometric properties of the translated and culturally adapted tests in German were in the same direction and of similar magnitude as the results from English-speaking countries [27]. An example of the benefits of cultural adaptation is evident from the lexical fluency test: it was well matched to US standards, indicating that the letters selected for the German language (letters D, S, and F) are an appropriate adaptation for word frequency, compared with the English language version (letters F, A, and S). Similar to US studies, the test-retest reliability was strong for most tests, and lower for other tests (e.g., CDT and BVMT-R copy) [28]: this is likely due to these tests being skewed toward maximum correct value. Strong alternate form reliability supports the use of a single set of norming tables to derive standardized scores. There was also indication of slight deviations in German language norms compared with English language norms, underscoring the importance of specific German language norms appropriate to the trial population. The use of these regional norms is crucial for reducing false-positive errors and costs of inaccurate case assignment that could occur if English norms were used. Taken together, current results suggest that the battery is fit for use in the German language, functioning in a similar manner to the English version.

Given the number of significant associations observed for age, age stratification is justified for generation of norming tables. Regarding gender effect for CVLT-II short-delay and long-delay, men consistently performed worse than women on these tests, a finding that has been reported in other studies using similar verbal memory measures [9,29]. The gender correction of +1 point for male participants could avoid the erroneous exclusion of cognitively healthy men from the study based on low, albeit normal, performance on the CVLT-II short-delay and long-delay subtests. We suggest that the future use of

the TOMMORROW battery in clinical practice may also benefit from gender corrections on these specific measures to facilitate accurate diagnostic inferences.

This study has several strengths. First, group classification of either NC or AD was based on an evaluation by a behavioral neurologist or a neuropsychologist, which was conducted independent of knowledge of the participant's performance on the TOMMORROW battery. Second, the sampling schema for the NC participants was uniquely designed to capture the population characteristics of individuals likely to participate in a preclinical AD trial, such as the TOMMORROW study. Sampling criteria that included gender and educational attainment for those aged 65–88 years provide a good representation of the range of normal performance in the local population. Third, this study had a robust basis for the normative values, which were based on several rigorous and unique study design elements. Translation of all measures was done in accordance with linguistic and cultural adaptation guidelines provided by International Society for Pharmacoeconomics and Outcomes Research [15]. Moreover, the same normative sample was co-normed providing uniformity in measurement and ready comparisons across tests and cognitive domains within participants, a capability that is unparalleled in current clinical practice given that these eight tests (CVLT-II, BVMT-R, Trails A and B, digit span, lexical fluency, semantic fluency, CDT, and MiNT) have never been co-normed together.

In conclusion, the German language TOMMORROW neuropsychological test battery is both linguistically and culturally valid and measures cognitive performance reliably across time. All tests were performed in a manner that closely parallels the English language standards. The few tests that differed slightly, compared with US normative data, nevertheless represent normal cognitive performance for this local population. The use of the neurocognitive battery and normative data derived from this study will improve precision for both measuring cognition in Switzerland and Germany and diagnosing incident MCI in clinical trials and clinical settings.

Acknowledgments

This work was sponsored by Takeda Development Center Americas, Inc., Deerfield, IL, USA; and Zinfandel Pharmaceuticals, Inc., Durham, NC, USA. Editorial support was provided by Chameleon Communications International Ltd, UK (a Healthcare Consultancy Group company) and sponsored by Takeda Pharmaceutical Company Ltd. The authors would also like to gratefully acknowledge the study participants and their project partners. In addition, they also gratefully acknowledge the clinical investigators and site staff at the Felix Platter Hospital in Basel, Switzerland.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.trci.2018.06.009.

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature using traditional sources and congress abstracts and presentations. Sensitive metrics are required for the detection of incident mild cognitive impairment due to Alzheimer's disease (MCI-AD). Composite end points based on selected neurocognitive measures are capable of reliably detecting cognitive decline in older adults at risk of developing AD. Relevant citations are provided.
2. Interpretation: In an international study, such as the TOMMORROW study that uses diagnostic criteria to detect MCI-AD, diagnostic standardization and validation of the metrics are essential to ensure that they capture the cognitive and functional features of early-stage AD across all the cultural settings included in the study. Here, we established the reliability and validity of the German language TOMMORROW neuropsychological test battery.
3. Future directions: The German language battery and normative data will improve precision for both measuring cognition and diagnosing incident MCI-AD in clinical trials and clinical settings in German-speaking countries.

References

[1] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984;141:1356–64.
[2] Reiman EM, Langbaum JB, Tariot PN, Lopera F, Bateman RJ, Morris JC, et al. CAP–advancing the evaluation of preclinical Alzheimer disease treatments. Nat Rev Neurol 2016;12:56–61.
[3] Langbaum JB, Hendrix SB, Ayutyanont N, Chen K, Fleisher AS, Shah RC, et al. An empirically derived composite cognitive test score with improved power to track and evaluate treatments for preclinical Alzheimer's disease. Alzheimers Dement 2014;10:666–74.
[4] Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol 2014; 71:961–70.
[5] Crenshaw DB, Gottschalk WK, Lutz MW, Grossman I, Saunders AM, Burke JR, et al. Using genetics to enable studies on the prevention of Alzheimer's disease. Clin Pharmacol Ther 2013;93:177–85.
[6] Welsh-Bohmer K, Romero H, Hayden K, Plassman B, Germain C, Sano M, et al. Challenges in international clinical trials to delay early symptomatic Alzheimer's disease. Alzheimers Dement 2013; 9:P137–8.
[7] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:270–9.

[8] Benedict RHB. Brief Visuospatial Memory Test Revised Professional Manual. Odessa, FL: Psychological Assessment Resources; 1997.
[9] Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test. 2nd ed. San Antonio, TX: Psychological Corporation; 2000.
[10] Reitan RM. The relation of the trail making test to organic brain damage. J Consult Psychol 1955;19:393–4.
[11] Gollan TH, Weissberger GH, Runnqvist E, Montoya RI, Cera CM. Self-ratings of spoken language dominance: a multi-lingual naming test (mint) and preliminary norms for young and aging spanish-english bilinguals. Biling (Camb Engl) 2012;15:594–615.
[12] Heaton RK. Revised Comprehensive Norms for an Expanded Halstead-Reitan Battery: Demographically Adjusted Neuropsychological Norms for African American and Caucasian Adults, Professional Manual. Lutz, FL: Psychological Assessment Resources; 2004.
[13] Wechsler D. WAIS-III WMS-III Technical Manual. San Antonio, TX: Psychological Corporation; 1977.
[14] Cahn DA, Salmon DP, Monsch AU, Butters N, Wiederholt WC, Corey-Bloom J, et al. Screening for dementia of the alzheimer type in the community: the utility of the Clock Drawing Test. Arch Clin Neuropsychol 1996;11:529–39.
[15] Wild D, Eremenco S, Mear I, Martin M, Houchin C, Gawlicki M, et al. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. Value Health 2009;12:430–40.
[16] 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.
[17] Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 1982;17:37–49.
[18] Walsh SP, Raman R, Jones KB, Aisen PS. ADCS prevention instrument project: the Mail-In Cognitive Function Screening Instrument (MCFSI). Alzheimer Dis Assoc Disord 2006;20:S170–8.
[19] Bridges AJ, Holler KA. How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. Child Neuropsychol 2007;13:528–38.
[20] Oosterhuis H, van der Ark LA, Sijtsma K. Sample size requirements for traditional and regression-based norms. Assessment 2015; 23:191–202.
[21] Albert MS, Moss MB, Tanzi R, Jones K. Preclinical prediction of AD using neuropsychological tests. J Int Neuropsychol Soc 2001;7:631–9.
[22] Backman L, Jones S, Berger AK, Laukka EJ, Small BJ. Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. Neuropsychology 2005;19:520–31.
[23] Welsh K, Butters N, Hughes J, Mohs R, Heyman A. Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. Arch Neurol 1991; 48:278–81.
[24] Chen P, Ratcliff G, Belle SH, Cauley JA, DeKosky ST, Ganguli M. Patterns of cognitive decline in presymptomatic Alzheimer disease: a prospective community study. Arch Gen Psychiatry 2001;58:853–8.
[25] Locascio JJ, Growdon JH, Corkin S. Cognitive test performance in detecting, staging, and tracking Alzheimer's disease. Arch Neurol 1995; 52:1087–99.
[26] Espeland MA, Vaughan L, Romero H, Hayden KM, Plassman BL, Welsh-Bohmer KA. Deriving an informed neurocognitive composite measure for delay-of-onset trials in mild cognitive impairment due to alzheimer's disease (mci-ad). Alzheimers Dement 2014;10:P776.
[27] Strauss E, Sherman EMS, Spreen O. A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. 3rd ed. Oxford, NY: Oxford University Press; 2006.
[28] Shulman KI. Clock-drawing: is it the ideal cognitive screening test? Int J Geriatr Psychiatry 2000;15:548–61.
[29] Welsh-Bohmer KA, Ostbye T, Sanders L, Pieper CF, Hayden KM, Tschanz JT, et al. Neuropsychological performance in advanced age: influences of demographic factors and Apolipoprotein E: findings from the Cache County Memory Study. Clin Neuropsychol 2009; 23:77–99.