Original Research

The Journal of Prevention of Alzheimer's Disease - JPAD
Volume 8, Number 2, 2021

Psychometric Properties of the Clinical Dementia Rating – Sum of Boxes and Other Cognitive and Functional Outcomes in a Prodromal Alzheimer's Disease Population

F. McDougall[1], C. Edgar[2], M. Mertes[3], P. Delmar[3], P. Fontoura[3], D. Abi-Saab[3], C.J. Lansdall[3], M. Boada[4,5], R. Doody[1,3]

1. Genentech, South San Francisco, USA; 2. Cogstate Ltd, London, UK; 3. F. Hoffmann-La Roche Ltd, Basel, Switzerland; 4. Research Center and Memory Clinic, Fundació ACE, Institut Català de Neurociències Aplicades, Universitat Internacional de Catalunya, Barcelona, Spain; 5. Networking Research Center on Neurodegenerative Diseases (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain.

Corresponding Author: Fiona McDougall, Genentech 620 E Grand Ave, South San Francisco, CA 94080, USA, mcdougall.fiona@gene.com

J Prev Alz Dis 2021;2(8):151-160
Published online December 21, 2020, http://dx.doi.org/10.14283/jpad.2020.73

Abstract

BACKGROUND: The Clinical Dementia Rating–Sum of Boxes (CDR-SB) has been proposed as a primary outcome for use in prodromal AD trials. However, the psychometric properties of this, and of other commonly used measures, have not been well-established in this patient population.

OBJECTIVE: To describe the psychometric properties of commonly used efficacy measures in a clinical trial of prodromal AD.

SETTING: Data were gathered as part of a two-year clinical trial.

PARTICIPANTS: Patients had biomarker confirmed prodromal AD.

MEASUREMENTS: Clinical Dementia Rating (CDR), Functional Activities Questionnaire (FAQ), Alzheimer's Disease Assessment Scale – Cognition Subscale 11 and 13 (ADAS-Cog), Mini Mental State Exam (MMSE), and Free and Cued Selective Reminding Test (FCSRT-IR [words]). Assessments were conducted at least every 24 weeks.

RESULTS: For the CDR-SB, test-retest reliability was good (intra-class correlation coefficient [ICC]=0.83); internal consistency was 0.65 at baseline but above 0.8 at later assessments. Relationships between the CDR-SB and other measures were as expected (higher correlations with more closely related constructs), and the CDR-SB differentiated between patients with different severities of dementia (-2.9 points difference between CDR-Global Score 0.5 and 1, P<.0001). Floor and ceiling effects on the CDR-SB total score were minimal; however, at baseline there were ceiling effects in the personal care domain. Further detail is provided on the psychometric properties of ADAS-Cog, MMSE, FCSRT-IR and FAQ in this population.

CONCLUSION: The psychometric properties of the CDR-SB are adequate in prodromal AD and continued use is warranted in clinical trials. However, there remains scope for improvement in the assessment of functional constructs and development of novel measures should continue.

Key words: Clinical dementia rating, prodromal Alzheimer's disease, psychometric testing.

Introduction

A number of clinical trials of potential disease-modifying treatments in Alzheimer's Disease (AD) are now targeted at early stage disease, where it is thought that there will be the greatest benefit to patients. By targeting the disease at the prodromal and mild dementia stages (also referred to in this paper as "early AD"), it is hoped to slow progression before extensive, irreversible neurodegeneration occurs. Since AD may be viewed as a continuum with preclinical, prodromal and dementia stages (mild, moderate and severe), the dementia diagnosis itself may be an important milestone in progression, but not one that represents a natural or stark differentiating boundary in terms of underlying pathophysiology. Diagnostic criteria for prodromal AD (pAD) (1) and mild cognitive impairment (MCI) due to AD (2) are now established; however, existing outcome measures to assess efficacy were mainly developed and validated for overt dementia and so may be unsuitable for clinical trials in this earlier patient population.

The FDA and EMA have both called for novel approaches to assess efficacy, recognizing the limitations of existing instruments in the earliest stages of AD. The FDA guidance and EMA guidelines (3, 4) have stated that clinical trials in the dementia stage of AD should use a co-primary approach, in which a treatment should demonstrate efficacy on both a cognitive measure and a functional measure (3, 5). This has been described as intending to ensure "that a clinically meaningful effect was established by a demonstration of benefit on the functional measure and that the observed functional benefit was accompanied by an effect on the core symptoms of the disease as measured by the cognitive assessment" (3). However, in the early stages of AD (stages 3 and 4), spanning pAD and mild AD dementia (mAD), it is recognized in both the FDA draft guidance (3) and the EMA guideline (6) that measurement may be more challenging. As independent research has shown,

Received May 5, 2020
Accepted for publication September 7, 2020

current assessment tools may have limited sensitivity due to ceiling effects and slow rates of progression (7, 8). Co-primary outcomes are not well established at this early stage, and whilst the principle behind the co-primary approach still holds, it has been suggested by regulators and others that application in practice could be achieved by integrated cognitive and functional endpoints, such as the Clinical Dementia Rating – Sum of Boxes (CDR-SB) score (6, 9-11). The CDR is intended to measure "the influence of cognitive loss on the ability to conduct everyday activities" (12). Whilst it has been hypothesized that the CDR may be broken down into 'cognitive' and 'functional' items (10, 13), the original intent was a unitary underlying construct (14) . Thus, it may be that observed statistical relationships supporting separate cognitive and functional items result from other properties, such as disease severity, or the use of information from the patient versus that from the caregiver-informant.

Studies have consistently reported high internal consistency for the CDR-SB across the AD spectrum, including clinically defined prodromal populations (CDR-GS = 0.5) (10, 15). Inter-rater reliability of the CDR-SB in a prodromal population is unclear, with some studies reporting low inter-rater agreement in populations with earlier non-biomarker confirmed AD dementia (13, 14, 16). Although many clinical trials in early/prodromal AD have used the CDR-SB as a primary endpoint, including studies of crenezumab (NCT02670083, NCT03114657), gantenerumab (NCT03443973, NCT03444870), aducanumab (NCT02484547, NCT02477800), BAN2401 (NCT03887455), and verubecestat (NCT01953601), a comprehensive assessment of the psychometric properties of the CDR-SB in this population is lacking.

To our knowledge, there are no studies describing the test-retest reliability according to gold-standard intra-class correlation coefficient (ICC) for a biomarker-confirmed prodromal population; a critical gap in the evidence needed to support use of the CDR-SB as a primary endpoint in AD clinical trials.

Investigation of the psychometric properties of commonly used outcome measures in the pAD clinical trials population is a critical step in confirming that assessments are fit for purpose, and for identifying potential gaps/areas for further development. Here, we describe traditional psychometrics, including test-retest reliability, of cognitive and functional assessments in a pAD trial population from SCarlet RoAD (NCT01224106; WN25203), a Phase 3, multicenter, randomized, double-blind, placebo-controlled study. In addition to amnestic MCI, subjects recruited to SCarlet RoAD were required to have evidence of amyloid pathology as demonstrated by low levels of Aβ(1–42) in cerebrospinal fluid (CSF). We also explore suitability of the CDR-SB as a single primary endpoint. Such properties should be established for the planned context of use (17, 18), and this paper is intended to do so for multinational, pAD, clinical trials. Furthermore, estimates are dataset dependent and a range of published estimates across different contexts of use may be informative. There are three main points of distinction from prior studies on this topic; some studies have used data from a single country only (15), while others have used observational cohorts (10) and defined AD based on clinical rather than biomarker criteria (10). This paper therefore adds to the existing literature by providing a comprehensive evaluation of the psychometric properties of key clinical outcome assessments in a multinational, clinical trial, biomarker confirmed prodromal AD population. The analysis presented is based on data from the SCarlet RoAD trial that evaluated low dose gantenerumab in patients with prodromal AD.

Methods

Data source and patients

The data were gathered as part of a Phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group, two-year study to evaluate the effect of subcutaneous gantenerumab (RO4909832) on cognition and function in prodromal Alzheimer's disease (pAD) conducted across 24 countries.

The primary objective of the trial was to evaluate the effect of gantenerumab versus placebo on the change from baseline to week 104 in the Clinical Dementia Rating scale Sum of Boxes (CDR-SB). All measures were translated and linguistically validated as per industry guidelines (19). CDR raters received comprehensive training prior to study start and refresher trainings at regular interval during the study. The information to make each rating was obtained through a semi-structured interview of the subject and a reliable informant. The study required that each subject have a study partner who, in the investigator's judgment, had frequent and sufficient contact with the subject so as to be able to provide accurate information regarding the subject's cognitive and functional abilities and who agreed to accompany the subject to clinic visits for scale completion. As far as possible, raters and study partners remained unchanged during the conduct of the study. Other assessments were rated by qualified site staff who were trained and, when necessary, certified to administer the assessments. Whenever possible, the CDR rater did not assess the other cognitive scales.

Inclusion criteria were modelled on International Working Group (IWG) criteria, which redefined AD as a clinicobiological syndrome that can be identified prior to the onset of dementia by an amnestic syndrome of the hippocampal type and supportive evidence from biomarkers (20). Key inclusion criteria were: age 50-85; recent gradual decline in memory (informant); abnormal

memory function based on the Free and Cued Selective Reminding Test (FCSRT: free recall <17, or total recall <40, or [free recall <20 and total recall <42]); Mini-Mental Status Exam (MMSE) score ≥24; and Clinical Dementia Rating - global score (CDR-GS) of 0.5 with memory box score of 0.5 or 1; CSF Aβ(1-42) ≤600 ng/L as measured by the central laboratory.

Study Design and Outcome measures

The study consisted of an 8-week screening period, a double-blind treatment phase of 100 weeks, a final assessment at week 104, and follow-up visits at 16 and 52 weeks after the last dose. Participants were recruited from clinical sites, some of which were memory centers. Subjects meeting all eligibility criteria during screening were randomized 1:1:1 to receive either placebo, 105 mg, or 225 mg gantenerumab subcutaneously every four weeks. Assessments at screening included the CDR, the Functional Activities Questionnaire (FAQ), the Alzheimer's Disease Assessment Scale – Cognition Subscale 11 and 13 item version (ADAS-Cog 11, ADAS-Cog 13 respectively), the MMSE, and the Free and Cued Selective Reminding Test (FCSRT-IR [words]) (Table 1). These assessments (listed in Table 1) were also conducted at baseline, and at 24-weeks interval (at weeks 24, 52, and 76) including final assessment at Week 104. The MMSE, ADAS-Cog, and FCSRT were obtained at additional 12 weeks intervals (at weeks 12, 36, 64, and 88), and the CDR was also obtained at weeks 64 and 88.

Clinical Dementia Rating scale

The CDR was originally developed as a staging tool to categorize dementia severity into normal, questionable, mild, moderate or severe. Clinicians rate the severity of symptoms across six domains following a semi-structured interview with the subject and a reliable informant or collateral source (e.g., family member). There are three cognition domains (Memory, Orientation, Judgment & Problem Solving) and three functional domains (Community Affairs, Home & Hobbies, and Personal Care) (12, 21). The response options for each domain describe five degrees of impairment: 0=None; 0.5=Questionable (not present in the Personal care domain); 1=Mild; 2=Moderate; 3=Severe. The CDR-Global score, which determines dementia stage, is rated from 0-3. The Sum of Boxes score is a continuous measure of dementia severity and ranges from 0-18. For completeness, we report the psychometric properties of the CDR-Cognition and CDR-Function domains individually, acknowledging that the CDR was not intended for this purpose and that each domain contains few items (3 items) for traditional analyses (e.g. Chronbach's α).

Alzheimer Disease Assessment Scale-Cognition (ADAS-Cog)

The ADAS-Cog-11 (22) is a structured scale that evaluates memory (word recall, word recognition), reasoning (following commands), language (naming, comprehension), orientation, ideational praxis (placing letter in envelope) and constructional praxis (copying geometric designs). Ratings of spoken language, language comprehension, word finding difficulty, and ability to remember test instructions are also obtained. The test is scored in terms of errors, with higher scores reflecting poorer performance. Scores can range from 0 (best) to 70 (worst). The 13-item version also includes Delayed Word Recall and Number Cancellation tasks, with scores ranging from 0 to 85.

Mini Mental State Exam (MMSE)

The MMSE consists of a set of standardized questions to evaluate possible cognitive impairment and help stage the severity level of this impairment. The questions target five areas; orientation, short term memory retention, attention, short term recall and language (23). The MMSE is scored as the number of correctly completed items with lower scores indicative of poorer performance and greater cognitive impairment. The total score ranges from 0 (worst) to 30 (best).

Functional Activities Questionnaire (FAQ)

The FAQ is an informant-based assessment in which caregivers rate abilities on 10 activities of daily living (ADLs) (24). The 10 items are scored as Dependent = 3, Requires assistance = 2, Has difficulty but does by self = 1, Normal = 0. The total score ranges from 0 to 30 with higher scores indicating worse functioning. The FAQ has demonstrated good sensitivity and specificity in differentiating MCI from very mild AD, by reflecting very mild functional impairment (25) .

Free and Cued Selective Reminding Test – Immediate Recall (FCSRT-IR)

The FCSRT-IR is a measure of memory under conditions that control attention and cognitive processing in order to obtain an assessment of memory unconfounded by normal age-related changes in cognition (26, 27). The FCSRT-IR used cards with four written words corresponding to a specific category cue, with immediate recall after each card followed by a cued recall using the category cue (28). Abnormal memory function according to FCSRT-IR was defined as a free recall score<17 (sum of free recall items), a total recall score<40 (sum of free recall and cued recall items), or a

Table 1. Clinical Outcome Assessments

Clinical Outcome Assessment (COA)	COA Type	Items	Domains Assessed	Range	Source	Interpretation
Clinical Dementia Rating						
CDR-Sum of Boxes CDR-Global Score	ClinRO	6	Functional impact of cognitive impairment: Memory, executive function, instrumental & basic activities of daily living	0-18 0-3	Interview with informant and patient, with PerfO elements	Higher scores represent greater severity of cognitive and functional impairment
Functional Activities Questionnaire						
	ClinRO	10	Instrumental activities of daily living	0-30	Interview with informant	Higher scores represent greater loss of independence in performing instrumental activities of daily living
Alzheimer's Disease Assessment Scale – Cognitive Subscale						
11-item Total Score	PerfO	11	Memory, language & praxis	0-72	Cognitive assessment of patient, with ClinRO elements	Higher scores represent greater severity of cognitive dysfunction
13-item Total Score		13	Additional memory & executive function items	0-87		
Mini-Mental State Exam						
	PerfO	11	Memory, language, praxis & executive function	0-30	Cognitive assessment of patient	Higher scores represent lower severity of cognitive dysfunction
Free and Cued Selective Reminding Test (FCSRT-IR [words])						
Free Recall Cued Recall Total Recall	PerfO	1	Learning & memory with controlled encoding	0-48	Cognitive assessment of patient	Higher scores represent better memory task performance

ClinRO: Clinician-rated Outcome Assessment; PerfO: Performance-based Outcome Assessment

free recall score<20 and total recall score<42. FCSRT-IR performance has been associated with preclinical and early dementia in several longitudinal epidemiological studies.

The CDR, ADAS-Cog, MMSE and FAQ may be viewed as composites, where a total score is based on the sum of item responses and individual items are intended to assess different cognitive and/or functional domains or concepts. Whilst total scores are also derived for FCSRT, items/words are not interpreted as individually meaningful.

Statistical methods

All screening and baseline analyses were conducted on the total sample. Analyses that included Week 52 and/or Week 104 data were conducted in the placebo group only to remove the potential impact of treatment from the evaluation of psychometrics. Patients were included in the analysis if they had completed measures at a given time point.

Test-retest reliability

Test–retest reliability is used to assess the degree to which a measure provides stable scores over time, assuming that the underlying condition of patients has not changed. This aspect of reliability was evaluated by intra-class correlation coefficients (ICCs, Shrout & Fleiss classification Random set 2, 1) (29) between the screening and baseline visits i.e. an interval approximately 8 Weeks (up to 12 Weeks was allowed for FCSRT-IR). Subjects were expected to remain clinically stable over this interval, whereas for longer intervals (baseline to Week 52, Week 52 to Week 104), clinical progression would be expected. Intra-class correlation coefficients that exceed 0.70 are generally assumed to be adequate (30).

Internal consistency

Internal consistency refers to the degree of association between the individual items that comprise a composite measure, and was measured by Cronbach's α, which generally increases as the inter-correlation amongst test items increases (31). As a general rule, >0.7 is considered an appropriate target for internal consistency (30, 32, 33). Internal consistency was not calculated for the FCSRT-IR outcomes since these are essentially single item constructs.

Construct validity

Construct validity refers to the extent to which a measure adequately assesses an intended concept and

Table 2. Clinical characteristics

	All Patients at baseline, n=797		Week 52 placebo group only, n=222*		Week 104 placebo group only, n=104†	
	Mean	SD	Mean	SD	Mean	SD
CDR-SB	2.11	0.97	2.67	1.61	3.04	2.13
CDR-Cognition	1.54	0.58	1.85	0.87	2.05	1.15
CDR-Function	0.57	0.57	0.82	0.86	0.99	1.12
FAQ	4.76	4.14	6.73	5.99	7.90	6.67
ADAS-Cog 11	13.82	5.21	15.45	7.44	16.67	7.61
ADAS-Cog 13	23.22	6.79	24.48	9.44	26.85	9.69
MMSE	25.68	2.2	24.63	3.70	23.44	3.77
FCSRT-IR Free	11.15	6.03	--	--	8.89	7.26
FCSRT-IR Cued	18.18	6.85	--	--	16.63	7.37
FCSRT-IR Total	29.33	10.69	--	--	25.52	12.13

*For FAQ, ADAS-Cog11, and ADAS-Cog13 n=221; †For FAQ, ADAS-Cog11, and ADAS-Cog13 n=105, FCSRT-IR assessments n=100; ADAS-Cog-11: the Alzheimer's Disease Assessment Scale – Cognition Subscale 11 item version, ADAS-Cog-13: the Alzheimer's Disease Assessment Scale – Cognition Subscale 13 item version, CDR-Cognition: Clinical Dementia Rating Cognition domain, CDR-Function: Clinical Dementia Rating Function domain, CDR-SB: Clinical Dementia Rating-Sum of Boxes, FAQ: the Functional Activities Questionnaire, FCSRT-IR: Free and Cued Selective Reminding Test – Immediate recall, MMSE: Mini Mental State Exam, SD: standard deviation.

may be evaluated by the association to other measures of both similar and different concepts. Relationships between the measures were examined in cross-section, using scores at baseline and change from baseline scores at Week 104. Spearman correlation coefficients (with Fisher's adjustment) were used to test the correlation between continuous variables. It was expected that objective cognitive measures would be inter-correlated (≥0.4), as would functional measures, but that correlation between cognition and function measures may be lower. The following thresholds were used to assess the strength of the relationship: <0.2: Weak, ≥0.2 to <0.4: Modest, ≥0.4 to <0.6: Moderate, ≥0.6 to 0.8: Strong; ≥0.8: Very strong (34, 35).

Ability to detect change (responsiveness to decline)

As AD is a progressive neurodegenerative disease, decline over time may be used as a way to assess ability to detect change. As an effect size metric, standardized response means (SRM) for the change from baseline in the placebo arm were calculated as SRM = mean change divided by the standard deviation of change, at Week 104. For convenience, we considered values ≥0.2 to <0.5 as low and ≥0.5 to <0.8 as moderate responsiveness (36) (Table 3). Ceiling and floor effects were determined according to the proportion that received the highest and lowest scores at baseline.

Known groups validity

The difference between CDR Global score = 0.5 (Questionable dementia) and CDR Global Score = 1 (Mild dementia) groups were calculated, for each of the variables. This evaluation was conducted at Week 52 (with the exception of FCSRT for which Week 104 was used as the only available time-point), since this maximized the sample size in both CDR-GS = 0.5 and CDR-GS = 1 groups; CDR-GS of 0.5 was an inclusion criterion at screening and a reduced sample size was available at Week 104. Independent samples t-tests were used to assess the statistical significance of the between groups differences. A significant difference between groups is generally considered to reflect reasonable known groups validity.

Results

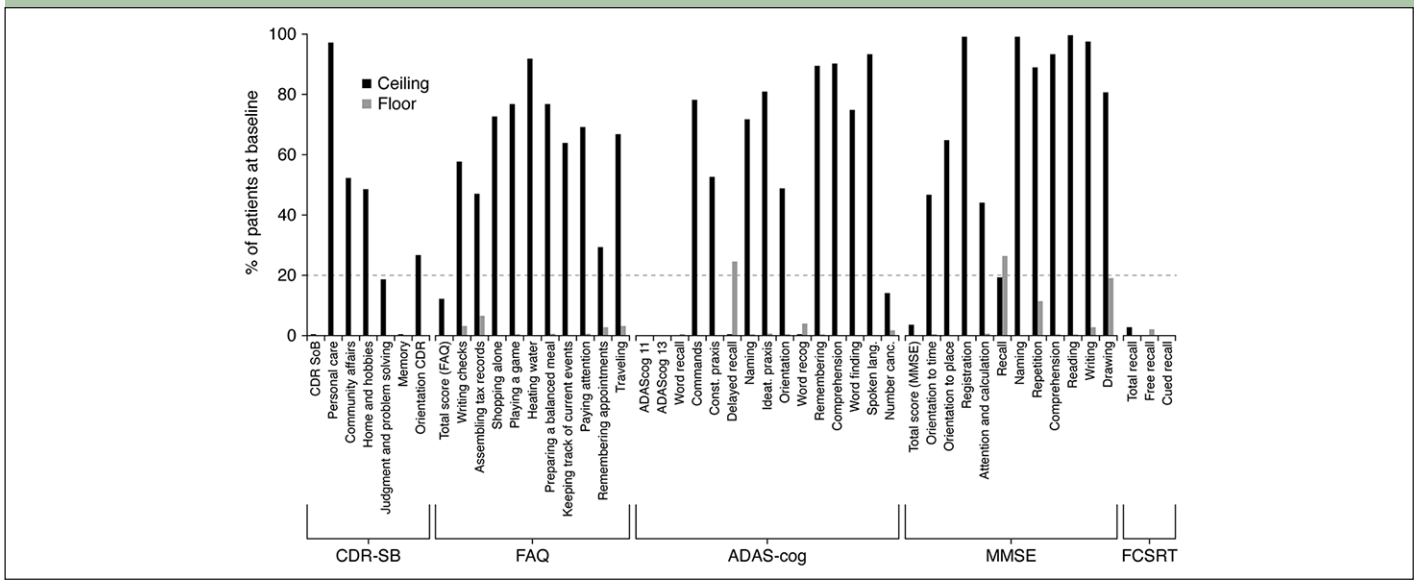
Patients

Seven hundred and ninety-seven subjects received allocated treatment (All Patients), mean age 70.4 years (SD 7.2), mean years of education 12.5 years (SD 4.5), 43.2% male. Two-hundred and sixty-six were randomized to placebo, with 104 completing the Week 104 visit (Placebo Arm), mean age 68.5 years (SD 6.8), mean years of education 12.3 years (SD 4.7), 43.8% male. The clinical characteristics of both populations are summarized in Table 2. Countries with the highest enrollment (>3%) included: the United States (14.3%), Spain (12.6%), Canada (7.3%), the United Kingdom (7.1%), Germany (6.9%), Italy (6.9%), France (6.4%), Australia (6.1%), Mexico (5.9%), Argentina (4.5%), and the Netherlands (4.1%).

Floor and Ceiling Effects

At the total score level, floor and ceiling effects were within acceptable ranges (Figure 1). At the item level, for all composite measures (i.e. CDR-SB, FAQ, ADAS-Cog and MMSE), notable ceiling effects (≥20% of patients at

Figure 1. Floor and Ceiling effects by item and total scores at baseline



All Patients. Dashed line represents threshold for notable floor or ceiling effect.

ceiling) were evident, showing that a large proportion of the enrolled pAD patient population were unimpaired in several of the items and/or domains assessed by these instruments (Figure 1). In addition, delayed word recall (ADAS-Cog and MMSE recall items) showed evidence of a floor effect.

Test-retest reliability

Test–retest reliability for the total scores was generally >0.7, with the exception of ADAS-Cog11 (0.67), MMSE (0.52) and FCSRT-IR Cued Recall (0.68) (Table 3).

Internal consistency

Internal consistency at screening and baseline was <0.5 for CDR-Cognition and CDR-Function, 0.65 for CDR-SB, 0.63 and 0.68 for ADAS-Cog11, and ADAS-Cog13, respectively; and 0.8 for FAQ. Chronbach's α tended to increase over the study, exceeding 0.7 for most measures at later timepoints (Table 3). The exception was the MMSE, which had very low internal consistency at baseline, rising to 0.66 at Week 104.

Construct validity

Inter-correlation of scores at baseline and change from baseline to Week 104 are reported in Table 4. CDR-SB was most strongly correlated with FAQ (0.6 at baseline and change from baseline). However, CDR-SB and FAQ were not strongly correlated with the cognitive measures, ADAS-Cog, MMSE or FCSRT (all correlations ≤0.4), with the exception of the correlation between change in CDR-SB and ADAS-Cog13 change at Week 104 (0.5). Both CDR 'cognition' and 'function' items were equally well correlated with function as measured by FAQ. However, a low degree of correlation was seen between CDR function and ADAS-Cog for the baseline scores only. As expected, the objective cognitive tests, ADAS-Cog, MMSE and FCSRT tended to be more highly correlated with each other.

Responsiveness to decline/Ability to detect change

Sensitivity to change was evaluated as the SRM (Table 3). Of the total scores, ADAS-Cog and FCSRT-IR were the least responsive, whilst CDR-SB, FAQ and MMSE were the most responsive. Importantly, CDR-SB and CDR cognition were free from floor and ceiling effects at Week 104, which may influence SRM. CDR function showed 25.6% at ceiling, FAQ 9.1% at ceiling and 0.3% at floor and MMSE 1.6% at ceiling, suggesting modest impact on SRM.

Known Groups Validity

For the evaluation of known groups validity, large and statistically significant (P<0.0001) differences were evident between subjects with a CDR-Global score of 0.5 versus those with a score of 1, for all measures (Table 3). CDR-SB, CDR-Function and CDR-Cognition all had Cohen's effect sizes of greater than 2 (≥small effect size).

Discussion

The Clinical Dementia Rating was devised as a global dementia-staging tool, taking into account results of clinician testing of cognitive performance and a rating of cognitive behavior in everyday activities, in six major

Table 3. Intraclass correlation coefficients, internal consistency, responsiveness and clinical validity

	Test-Retest ICC		Internal Consistency α		Responsiveness SRM		Know Groups Validity Mean diff between CDR Global 0.5 and 1*	
	Baseline to screening n=797	Baseline, n=797	Week 52, n=222	Week 104, n=104	Week 52 n=222	Week 104 n=104	Total score difference Week 52 n=222	Cohen's effect size Week 52 n=222
CDR-SB	0.83	0.65	0.84	0.90	0.55	0.71	−2.94	2.88
CDR-Cognition	0.76	0.34	0.63	0.72	0.51	0.74	−1.42	2.16
CDR-Function	0.8	0.18	0.54	0.69	0.42	0.55	−1.52	2.65
FAQ	0.82	0.8	0.88	0.92	0.54	0.73	−8.83	1.88
ADAS-Cog 11	0.67	0.63	0.77	0.82	0.31	0.55	−6.76	1.18
ADAS-Cog 13	0.72	0.68	0.77	0.83	0.22	0.59	−8.8	1.16
MMSE	0.52	0.17	0.6	0.66	-0.39^	-.71^	3.31	1.08
FCSRT-IR Free	0.75	-	-	-		-0.50^	5.99	0.88
FCSRT-IR Cued	0.68	-	-	-		-0.20^	5.38	0.76
FCSRT-IR Total	0.82	-	-	-		-0.46^	11.37	0.98

*All p values less than .0001; evaluation was conducted at Week 52 for all measures, with the exception of FCSRT for which Week 104 was used, in order to maximize n; ^Negative value due to scoring direction (lower score = worse cognition); α, standardized Cronbach's alpha; ADAS-Cog-11: the Alzheimer's Disease Assessment Scale – Cognition Subscale 11 item version, ADAS-Cog-13: the Alzheimer's Disease Assessment Scale – Cognition Subscale 13 item version, CDR-Cognition: Clinical Dementia Rating Cognition domain, CDR-Function: Clinical Dementia Rating Function domain, CDR-SB: Clinical Dementia Rating-Sum of Boxes, FAQ: the Functional Activities Questionnaire, FCSRT-IR: Free and Cued Selective Reminding Test – Immediate recall, MMSE: Mini Mental State Exam. Internal consistency not appropriate for FCSRT.

Table 4. Inter-correlation of scores

	Baseline (n=797)										Change from baseline to Week 104 (n=104)									
	CDR-SB	CDR-Cognition	CDR-Function	FAQ	ADAS-Cog 11	ADAS-Cog 13	MMSE	FCSRT-IR Free	FCSRT-IR Cued	FCSRT-IR Total	CDR-SB	CDR-Cognition	CDR-Function	FAQ	ADAS-Cog 11	ADAS-Cog 13	MMSE	FCSRT-IR Free	FCSRT-IR Cued	FCSRT-IR Total
CDR-SB	-	0.9	0.8	0.6	0.3	0.3	−0.2	−0.3	−0.2	−0.3	-	0.9	0.9	0.6	0.4	0.5	−0.4	−0.3	−0.2	−0.3
CDR-Cognition		-	0.5	0.5	0.4	0.4	−0.3	−0.3	−0.2	−0.3		-	0.6	0.5	0.4	0.4	−0.4	−0.3	−0.2	−0.3
CDR-Function			-	0.5	0.2	0.2	−0.1	−0.2	−0.1	−0.2			-	0.5	0.4	0.4	−0.3	−0.3	−0.1	−0.3
FAQ				-	0.2	0.2	−0.2	−0.2	−0.1	−0.2				-	0.3	0.3	−0.4	−0.2	−0.2	−0.2
ADAS-Cog 11					-	0.9	−0.5	−0.5	−0.4	−0.5					-	0.9	−0.5	−0.3	−0.2	−0.4
ADAS-Cog 13						-	−0.5	−0.6	−0.4	−0.5						-	−0.5	−0.3	−0.2	−0.4
MMSE							-	0.3	0.2	0.3							-	0.3	0.2	0.3
FCSRT-IR Free								-	0.4	0.8								-	0.1	0.6
FCSRT-IR Cued									-	0.8									-	0.8
FCSRT-IR Total										-										-

ADAS-Cog-11: the Alzheimer's Disease Assessment Scale – Cognition Subscale 11 item version, ADAS-Cog-13: the Alzheimer's Disease Assessment Scale – Cognition Subscale 13 item version, CDR-Cognition: Clinical Dementia Rating Cognition domain, CDR-Function: Clinical Dementia Rating Function domain, CDR-SB: Clinical Dementia Rating-Sum of Boxes, FAQ: the Functional Activities Questionnaire, FCSRT-IR: Free and Cued Selective Reminding Test – Immediate recall, MMSE: Mini Mental State Exam.

categories of cognitive performance. Impairment is scored as decline from the person's previous level due to cognitive loss alone, not impairment due to other factors, such as physical impairment, depression, or personality change (21). The CDR is considered to be a face valid measure of "the influence of cognitive loss on the ability to conduct everyday activities" (12). The CDR-SB has gained prominence in recent times as a single primary endpoint for clinical trials in early AD. Results from the crenezumab discontinued Ph III trial show a robust decline in CDR-SB score over 24 months in patients with both prodromal and mild AD, suggesting that, when enriching for "fast progressors" who are impaired on the FCSRT , the CDR-SB is sensitive to decline in early (prodromal-mild) AD (37). Whilst psychometric properties of the CDR-SB have been explored in early AD dementia (10), they have not been explored in a biomarker confirmed pAD population, and test-retest reliability, critical to repeated assessments, has not previously been published to our knowledge.

These data further support the psychometric properties of the CDR-SB, in demonstrating adequate test-retest reliability, a good degree of internal consistency especially over time, and also construct validity in terms of association to instrumental activities of daily living measured by the FAQ. Importantly, these properties

are now confirmed in a pAD clinical trial population. Although there was evidence for ceiling effects in individual domains/items at the baseline assessment, this did not have a major impact on sensitivity to decline, and CDR-SB showed a greater degree of responsivity than ADAS-Cog and FCSRT-IR. However, SRM was lower than previously reported over two years in an early AD population derived from ADNI data (0.71 in this report, versus 1.03 in ADNI), which may result from differences in inclusion criteria (10). There were floor and ceiling effects at the item level for all composite measures; this may be an important consideration with respect to the coverage of relevant concepts and for potential sensitivity to disease progression in the early stages of the disease. Floor effects suggest that even at the early stages of disease, delayed free recall assessments may be markedly impaired, again impacting potential sensitivity.

Previous reports have focused on inter and intra-rater reliability. The novel finding for test-retest is of particular value, given the importance of reliability in clinical trial use. Strong test-retest reduces measurement error, which increases the likelihood of detecting true treatment effects. There was an improvement in internal consistency from baseline for all measures over the course of the study. One possible explanation for improved internal consistency may be 'other' reliability, such as improved intra-rater reliability and reliability of subject and informant report, as all parties become more familiar with the scales and have more data available to inform them. In addition, regression to the mean, or disease progression bias could result in greater homogeneity of scores between items over time. The internal consistency for CDR-SB at Week 52 and 104 (Cronbach's alpha 0.84 and 0.90, respectively) was similar to that observed in the French REAL.FR cohort study of patients with very mild-to-moderate AD (0.88) (15).

Specific to construct validity of the CDR, Tractenberg et al previously observed that in an AD dementia population, change in 'cognitive' items showed a modest correlation with change in MMSE and a low correlation with change in ADL, and 'functional' items the opposite pattern (13). Along with results from principal components analyses, this was seen as supportive of separate cognitive and function domains. In the present data, a correlation was observed between both the CDR cognition and function domains and FAQ, for both baseline and change scores (all 0.5). This may be seen as supportive of overall convergent validity with the FAQ (25). Inter-correlation of CDR-SB and FAQ may be driven by measurement of function and some direct overlap in item content and the use of informant report in both assessments. Cedarbaum et al (2013), found correlations with FAQ tended to be higher than with ADAS-Cog11 or ADAS-Cog13 for both cognitive (0.63, 0.55, and 0.59, respectively) and function domains (0.58, 0.42, and 0.45, respectively) in subjects with early or mild AD at baseline in the ADNI study (10). In addition, although their factor analysis showed some support for separate domains, there was overlap for "Judgment and problem solving" and "Community affairs" items in several cases, and a differential pattern based on disease severity was observed. Thus, the CDR may not capture function and cognition as separate domains but still address both, consistent with the original unitary measurement concept ("the influence of cognitive loss on the ability to conduct everyday activities"). Furthermore, low internal consistency reliability of these scores suggests there may be too few items for them to be reliable as separate measures.

For the FAQ and ADAS-Cog, adequate test-retest reliability and a good degree of internal consistency were observed. Both measures demonstrated construct validity in terms of association to related measures (CDR to FAQ and ADAS-Cog to MMSE and FCSRT). This is an important finding for the FAQ, given the lack of validation evidence for the scale beyond discriminative ability (38). Though there was evidence for ceiling effects with individual items at the baseline assessment for both measures, this did not have a major impact on sensitivity to decline for FAQ as the SRM for FAQ (0.73) was greater than for the other measures.

For the MMSE, assessment of psychometric properties was impacted by its use as a screening criterion, with only scores of between 24 and 30 out of 30 possible at screening. This would initially reduce range of scores and variance, decreasing power to obtain high alpha coefficients and impact the ability to adequately assess scale properties at the screening and baseline assessments in particular. Thus, caution is warranted in the interpretation of the results. Overall MMSE did show good sensitivity to decline, comparable to CDR-SB and FAQ (SRM=−0.71), with orientation to time as the single greatest contributory item (SRM=−0.63). This prominence of orientation as sensitive to decline across the different scales is consistent with other data, which has shown orientation to be sensitive to disease progression and important for inclusion in novel composite outcomes (39).

For the FCSRT-IR, adequate test-retest reliability was also observed, though this was also utilized as a screening inclusion criterion. Whilst the measures of free, cued, and total recall were relatively free from ceiling and floor effects, this did not translate to sensitivity to decline and SRMs were −0.2 for cued, −0.5 for free and −0.46 for total recall. Therefore, there may be limited additional value in FCSRT-IR as a longitudinal outcome measure in this patient population.

There are some limitations which could impact the generalizability of these findings. The study population was derived from a clinical trial, in which CDR-Global Score was one of the inclusion criteria, and thus we cannot rule out the possibility that this influenced the reporting of the CDR domains at baseline. Additionally, the CDR-Global score was used to define questionable dementia and mild dementia for the known groups

validity analysis of CDR-SB. This may have impacted the analysis, as the CDR-SB and global score may be interrelated. Although industry standards were followed with regards to translation, we did not formally evaluate whether psychometric properties differed by culture or language. Furthermore, this was a biologically homogenous population with low levels of Aβ(1-42) and different results may be found in a more heterogeneous sample. This study population may have been subject to selection bias due to initial study recruitment methods (e.g. site selection), as well as individual interest in participating in a clinical trial. Loss to follow-up will also have impacted the representativeness of longitudinal analyses. Finally, further work is required to establish what constitutes a meaningful change on the CDR-SB in prodromal AD.

In conclusion, CDR-SB showed adequate psychometric properties in the pAD population and its sensitivity to decline over time further support its utility as a clinical trial outcome measure. In addition, its conceptual basis as a measure of the influence of cognitive loss on the ability to conduct everyday activities was supported by the construct validity data. These data reinforce the continued use of CDR-SB as a single primary outcome measure in early AD clinical trials, such as the phase III gantenerumab GRADUATE program and the phase III BAN2401 Clarity AD trial. In addition, validity and reliability of the other assessments, particularly the FAQ, is further supported.

Efforts to develop novel cognitive and functional assessments free from ceiling and floor effects and with greater sensitivity to change in this population should continue. However, given the adequate psychometrics of the CDR-SB, its clinical relevance, and the lack of a clearly established relationship between other objective cognitive endpoints and clinical benefit, at least for patients with early AD, there is good reason to continue to employ the CDR-SB in treatment trials.

Acknowledgements: The authors would like to acknowledge to important contributions of the SCarlet RoAD Investigators, Patients and their Families participating globally in Argentina, Australia, Belgium, Brazil, Canada, Chile, the Czech Republic, Denmark, Finland, France, Germany, Italy, Mexico, the Netherlands, Poland, Portugal, Russia, Spain, South Korea, Sweden, Switzerland, Turkey, the United Kingdom and the United States. This work was supported by F. Hoffmann-La Roche Ltd, Basel, Switzerland.

Conflicts of interest: FM is an employee of Genentech Inc., South San Francisco, USA. MM, PD, PF, DAS, and CJL are employees of F. Hoffmann-La Roche Ltd, Basel, Switzerland. PD owns stock in F. Hoffmann-La Roche Ltd. RD is an employee and owns stock in Genentech Inc. and F. Hoffmann-La Roche Ltd. MB has received grants from Merck & Co., Inc. related to the submitted work (paid to the institution); she has received grants from Araclon, Biogen Research Ltd, Bioberica, Grifols, Lilly S.A, Merck Sharp & Dohme, Nutricia SRL, Oryzon Genomics, Piramal Imaging Ltd, Schwabe Farma Iberica SLU and Merck & Co, Inc. within the last 36 months outside the submitted work (paid to the institution); she has served as a consultant or provided scientific advisory board services and/or given lectures for Roche, Araclon, Bioberica, Grifols, Kyowa Hakko Kirin, Laboratorios Servier, Lilly, S.A., Merck Sharp & Dohme, Nutricia SRL, Schwabe Farma Iberica SLU.

Ethical Standards: Institutional Review Boards (IRBs) approved the SCarlet RoAD study, and all participants gave informed consent before participating.

Data sharing statement: Qualified researchers may request access to individual patient-level data through the clinical study data request platform: https://vivli.org. Further details on Roche's criteria for eligible studies are available here: https://vivli.org/members/ourmembers. For further details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here: https://www.roche.com/research_and_development/who_we_are_how_we_work/clinical_trials/our_commitment_to_data_sharing.htm

References

1. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol. 2014;13(6):614-29.
2. 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(3):270-9.
3. FDA. Early Alzheimer's Disease: Developing Drugs for Treatment Guidance for Industry, Draft Guidance. 2018.
4. EMA. Discussion paper on the clinical investigation of medicines for the treatment of Alzheimer´s disease and other dementias. In: CHMP, editor. London: European Medicines Agency; 2014. p. 33.
5. FDA. Multiple Endpoints in Clinical Trials. In: (CDER) CfDEaR, (CBER) CfBEaR, editors. Rockville, MD2017.
6. EMA. Guideline on the clinical investigation of medicines for the treatment of Alzheimer's disease In: CHMP, editor. London2018.
7. Jekel K, Damian M, Wattmo C, Hausner L, Bullock R, Connelly PJ, et al. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. Alzheimers Res Ther. 2015;7(1):17.
8. Wang J, Logovinsky V, Hendrix SB, Stanworth SH, Perdomo C, Xu L, et al. ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. J Neurol Neurosurg Psychiatry. 2016.
9. FDA. Guidance for Industry Alzheimer's Disease: Developing Drugs for the Treatment of Early Stage Disease DRAFT GUIDANCE. In: CDER, editor. 2013.
10. Cedarbaum JM, Jaros M, Hernandez C, Coley N, Andrieu S, Grundman M, et al. Rationale for use of the Clinical Dementia Rating Sum of Boxes as a primary outcome measure for Alzheimer's disease clinical trials. Alzheimers Dement. 2013;9(1 Suppl):S45-55.
11. Kozauer N, Katz R. Regulation of drugs for early Alzheimer's disease. N Engl J Med. 2013;369(3):288.
12. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993;43(11):2412-4.
13. Tractenberg RE, Weiner MF, Cummings JL, Patterson MB, Thal LJ. Independence of changes in behavior from cognition and function in community-dwelling persons with Alzheimer's disease: a factor analytic approach. J Neuropsychiatry Clin Neurosci. 2005;17(1):51-60.
14. Morris JC, Ernesto C, Schafer K, Coats M, Leon S, Sano M, et al. Clinical dementia rating training and reliability in multicenter studies: the Alzheimer's Disease Cooperative Study experience. Neurology. 1997;48(6):1508-10.
15. Coley N, Andrieu S, Jaros M, Weiner M, Cedarbaum J, Vellas B. Suitability of the Clinical Dementia Rating-Sum of Boxes as a single primary endpoint for Alzheimer's disease trials. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2011;7(6):602-10.e2.
16. Rockwood K, Strang D, MacKnight C, Downer R, Morris JC. Interrater reliability of the Clinical Dementia Rating in a multicenter trial. J Am Geriatr Soc. 2000;48(5):558-9.
17. FDA. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims:Guidance for Industry. 2009.
18. Powers JH, 3rd, Patrick DL, Walton MK, Marquis P, Cano S, Hobart J, et al. Clinician-Reported Outcome Assessments of Treatment Benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. Value Health. 2017;20(1):2-14.
19. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation. Value in Health. 2005;8(2):94-104.
20. Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 2007;6(8):734-46.
21. Berg L. Clinical Dementia Rating (CDR). Psychopharmacol Bull. 1988;24(4):637-9.
22. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry. 1984;141(11):1356-64.
23. Folstein MF, Folstein SE, McHugh PR. «Mini-mental state». A practical method for grading the cognitive state of patients for the clinician. Journal of

psychiatric research. 1975;12(3):189-98.
24. Vellas B, Andrieu S, Sampaio C, Coley N, Wilcock G, European Task Force G. Endpoints for trials in Alzheimer's disease: a European task force consensus. Lancet Neurol. 2008;7(5):436-50.
25. Pfeffer RI, Kurosaki TT, Harrah CH, Jr., Chance JM, Filos S. Measurement of functional activities in older adults in the community. J Gerontol. 1982;37(3):323-9.
26. Grober E, Buschke H. Genuine memory deficits in dementia. Developmental Neuropsychology. 1987;3:13-36.
27. Grober E, Hall C, McGinn M, Nicholls T, Stanford S, Ehrlich A, et al. Neuropsychological strategies for detecting early dementia. J Int Neuropsychol Soc. 2008;14(1):130-42.
28. Grober E, Sanders AE, Hall C, Lipton RB. Free and cued selective reminding identifies very mild dementia in primary care. Alzheimer Disease and Associated Disorders. 2010;24(3):284.
29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.
30. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd. ed. ed. New York: McGraw-Hill; 1994.
31. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
32. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. Journal of Nursing scholarship. 2007;39(2):155-64.
33. DeVellis RF. Scale development: Theory and applications: Sage publications; 2016.
34. Swinscow T. Correlation and regression. Statistics at square one. 9 ed1997.
35. Bartz AE. Basic Statistical Concepts. 4 ed. Upper Saddle River, NJ: Merrill; 1999.
36. Middel B, Van Sonderen E. Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. International Journal of Integrated Care. 2002;2.
37. Sink K, Djakovic S., Smith JW, Hu N, Mackey H, Ostrowitzki S, et al. FCSRT inclusion criteria support recruitment of a population with early Alzheimer's disease likely to progress over 24 months: results from the CREAD trial. Clinical Trials in Alzheimer's Disease; San Diego, USA2019.
38. Kaur N, Belchior P, Gelinas I, Bier N. Critical appraisal of questionnaires to assess functional impairment in individuals with mild cognitive impairment. Int Psychogeriatr. 2016;28(9):1425-39.
39. Vellas B, Bateman R, Blennow K, Frisoni G, Johnson K, Katz R, et al. Endpoints for Pre-Dementia AD Trials: A Report from the EU/US/CTAD Task Force. J Prev Alzheimers Dis. 2015;2(2):128-35.