doi:10.1093/brain/awn234

Development of methodology for conducting clinical trials in frontotemporal lobar degeneration

David S. Knopman,[1] Joel H. Kramer,[2] Bradley F. Boeve,[1] Richard J. Caselli,[3] Neill R. Graff-Radford,[4] Mario F. Mendez,[5] Bruce L. Miller[2] and Nathaniel Mercaldo[6]

[1]Mayo Clinic, Rochester, MN, [2]University of California, San Francisco, San Francisco, CA, [3]Mayo Clinic, Scottsdale, AZ, [4]Mayo Clinic, Jacksonville, FL, [5]University of California Los Angeles, Los Angeles, CA and [6]University of Washington, Seattle, WA, USA

Correspondence to: David S. Knopman, MD, Department of Neurology, Mayo Clinic College of Medicine, 200 First Street SW, Rochester MN 55905, USA
E-mail: knopman@mayo.edu

To design clinical trials for the frontotemporal lobar degenerations (FTLD), knowledge about measurement of disease progression is needed to estimate power and enable the choice of optimal outcome measures. The aim here was to conduct a multicentre, 1 year replica of a clinical trial in patients with one of four FTLD syndromes, behavioural variant frontotemporal dementia (bvFTD), progressive nonfluent aphasia (PNFA), progressive logopenic aphasia (PLA) and semantic dementia (SMD). Patients with one of the four FTLD syndromes were recruited from five academic medical centres over a 2 year period. Standard operationalized diagnostic criteria were used. In addition to clinical inclusion and exclusion criteria, patients were required to exhibit focal frontal, temporal or insular brain atrophy or dysfunction by neuroimaging. Patients underwent neuropsychological, functional, behavioural, neurological and MR imaging assessment at baseline and approximately 12 months later. Potential outcome measures were examined for their rates of floor and ceiling values at baseline and end of study, their mean changes and variances. The neuropsychological tests were combined into two cognitive composites—one for language functions and the other for executive functions. There were 107 patients who underwent baseline assessment and 78 who completed a follow-up assessment within 10–16 months. Two global measures, the FTLD-modified Clinical Dementia Rating (FTLD-modified CDR) and the Clinical Global Impression of Change (CGIC) demonstrated decline in the majority of patients. Several cognitive measures showed negligible floor or ceiling scores either at baseline or follow-up. Scores declined at follow-up in the majority of patients. The cognitive, executive and combined composites were shown to be sensitive to change across all FTLD syndromes. Patients improved at follow-up on the behavioural scales—the Frontal Behavioural Inventory (22%) and the Neuropsychiatric Inventory (28%)—suggesting that these instruments may not be ideal for clinical trial use. It was feasible to recruit FTLD patients in a simulated multi-centre trial. There are several candidate outcome measures—including the FTLD-CDR and the cognitive composites—that could be used in clinical trials across the spectrum of FTLD.

Keywords: frontotemporal dementia; clinical trials; neuropsychology

Abbreviations: bvFTD = behavioural variant frontotemporal dementia; CGIC = Clinical Global Impression of Change; FAQ = Functional Activities Questionnaire; FBI = Frontal Behavioral Inventory; FTLD = frontotemporal lobar degenerations; FTLD-CDR = FTLD-specific Clinical Dementia Rating; MMSE = Mini-Mental State Examination; NPI = Neuropsychiatric Inventory; PLA = progressive logopenic aphasia; PNFA = progressive nonfluent aphasia; SMD = semantic dementia

Received March 17, 2008. Revised August 27, 2008. Accepted August 28, 2008. Advance Access publication October 1, 2008

Introduction

The frontotemporal lobar degenerations (FTLD) represent a challenge for therapeutic development because of their clinical and pathological heterogeneity. Depending upon the location of the major pathological changes, either language or behavioural symptoms may predominate in the clinical presentation. Four syndromes have been described that represent one of the FTLD's pathologically: behavioural variant frontotemporal dementia (bvFTD), progressive nonfluent aphasia (PNFA), semantic dementia (SMD) and progressive logopenic aphasia (PLA). The first three are highly likely to represent FTLD pathologically, while PLA may sometimes represent an FTLD, but also sometimes Alzheimer's disease (Galton *et al.*, 2000; Gorno-Tempini *et al.*, 2004; Kertesz *et al.*, 2005; Knibb *et al.*, 2006; Josephs *et al.*, 2008; Mesulam *et al.*, 2008). Future biomarkers may allow greater predictive accuracy for diagnoses based on underlying molecular biology, but at the present time, clinical diagnoses are the best means of classifying FTLD patients.

Although there are currently no treatments for the FTLD's, understanding of basic biology of the FTLD's may soon lead to promising compounds with therapeutic potential. As of 2008, the few clinical trials involving FTLD subjects have been small, single-centre studies, and only some were double-blind (Boxer and Boeve, 2007). The prior trials generally used rating scales for psychiatric symptoms. Cognitive measures were rarely used (Deakin *et al.*, 2004). In addition to limited experience with cognitive, global and functional assessments in FTLD clinical trials, there is also very limited information about longitudinal change on potential trial instruments in patients with FTLD (Rascovsky *et al.*, 2005; Wicklund *et al.*, 2007; Kipps *et al.*, 2008). Quantitative, longitudinal information on the outcome measures is required to estimate power and sample sizes.

In order to develop a framework for multicentre, randomized clinical trials of new agents for the FTLD spectrum, we conducted a replica of a clinical trial by recruiting patients with FTLD and followed them for 1 year. The purpose of this study was to conduct a multicentre trial in patients with bvFTD, PNFA, SMD and PLA. In the current paper, we discuss the cognitive, functional and behavioural assessments. All patients also underwent serial MR imaging, but the imaging data will be reported separately.

The current study addressed several important questions about instruments for FTLD clinical trials. What assessment instruments are best to detect disease progression? What tests have the lowest rates of ceiling and floor values? Which tests have the largest longitudinal change and the smallest variation? Ideally, the ratio of change over time to variability should be as large as possible to enhance power. What instruments could serve as primary outcome measures? How many patients will be needed to avoid a type 2 error? Answers to these questions should be available prior to designing and conducting a clinical trial in FTLD.

Methods

Patients with mild to moderate cognitive impairment due to one of the four syndromes—bvFTD, PNFA, PLA and SMD—were recruited from five academic medical centres over a 2 year period. Standard diagnostic criteria were operationalized for clinical trial usage. In order to replicate an actual clinical trial, all patients were examined at baseline and then re-evaluated 12 months later. The study was approved by the Institutional Review Boards at all five sites.

Diagnostic criteria

The inclusion criteria for this study were based on the 'Neary' criteria (Neary *et al.*, 1998), and the aggregate expert opinion of the investigators involved in this project (Gorno-Tempini *et al.*, 2004; Josephs *et al.*, 2006). The criteria focussed on a predominance of a frontal or temporal lobe cognitive/behavioural syndrome and the absence or insignificance of an anterograde amnesia and visuospatial impairment in the initial clinical presentations (i.e. within the first 2 years of symptoms). In addition, all patients were required to have imaging studies demonstrating focal cerebral atrophy of at least one of the following: the anterior temporal lobes, frontal lobes and insula or caudate nuclei. The criteria have been previously reported (Knopman *et al.*, 2007).

PNFA

Inclusion criteria were: A 6-month history of difficulty with expressive speech characterized by at least three of the following: nonfluency (reduced numbers of words per utterance), speech hesitancy or laboured speech and word finding difficulty or agrammatism, where these symptoms constitute the principal deficits and the initial presentation.

PLA

Inclusion criteria were: A 6-month history of fluent aphasia with anomia but intact word meaning and object recognition, where these symptoms constitute the principal deficits and the initial presentation.

SMD

Inclusion criteria were: A 6-month history of loss of comprehension of word meaning, object identity or face identity, where these symptoms constitute the principal deficits and the initial presentation.

bvFTD

Inclusion criteria were: A 6-month history of change in personality and behaviour sufficient to interfere with work or interpersonal relationships, these symptoms constituted the principal deficits and the initial presentation, and with at least five of the following: (i) Loss of insight; (ii) Reduced empathy; (iii) Disinhibition; (iv) Impulsivity; (v) Apathy; (vi) Social withdrawal and disengagement; (vii) Restlessness; (viii) Poor self-care; (ix) Emotional lability; (x) Appetite disturbance or hyperorality; (xi) Easily distracted; and (xii) Compulsive or stereotypic behaviours.

The exclusion criteria (Knopman *et al.*, 2007) were common to all four syndromes and included the following:

(i) Expressive language deficits such that the patient was too severely impaired to allow testing at baseline. As a guideline, an orientation score of ⩾6 on the Mini-Mental State Examination (MMSE) was used (Folstein *et al.*, 1975).
(ii) The syndrome is due to cerebrovascular disease.
(iii) The syndrome is due to traumatic brain injury.
(iv) By clinical history or neuropsychological testing, anterograde amnesia is a principal symptom or sign.
(v) By clinical history or neuropsychological testing, visuospatial deficits are principal symptoms/signs.
(vi) Rapid eye movement sleep behaviour disorder is present.

(vii) Imaging findings are diagnostic of another neurological disease, including the presence of >1 lacunar infarction.
(viii) Unable to undergo MR at baseline.
(ix) Severity of symptoms such that patient is not expected to be able to undergo MR imaging 1 year later.
(x) Prior history dating from early adulthood of schizophrenia, bipolar disease, mental retardation and severe personality disorder.
(xi) The clear, unequivocal presence of another neurological disease such as Parkinson's disease, Huntington's disease, progressive supranuclear palsy, multiple sclerosis; an inherited disorders such as metachromatic leukodystrophy, or any other defined neurological disorder other than an FTLD.
(xii) Not a native English speaker premorbidly.
(xiii) There is no caregiver who sees the patient at least once per week.

Procedures

Prior to the initiation of recruitment, all investigators, study coordinators and psychometricians met and reviewed procedures. The investigators reached a working consensus on common definitions of the diagnostic criteria and procedures.

At the initial study visit, patients were evaluated for participation based on inclusion and exclusion criteria. If judged to have capacity to do so by the site principal investigator, written informed consent was obtained from the patient. Otherwise, informed consent was obtained from the next of kin caregiver. Demographic information was obtained from the caregiver and the patient. Inclusion and exclusion criteria were verified. After that, the patient underwent cognitive assessment, the patient's caregiver was queried using several instruments, and a syndromic diagnosis was made. An MR scan was obtained.

Patients were re-evaluated 12 months later using the same procedures.

Instruments

Global assessment of change

In keeping with the standards set in Alzheimer's disease therapeutic trials, there is a need for a global assessment tool for FTLD. The value of a global rating is that it conveys a sense of change based on the judgment of a clinician who has weighed and integrated information from both informant interviews and patient examinations. A modified Alzheimer's Disease Cooperative Study Clinical Global Impression of Change (ADCS-CGIC) (Schneider *et al.*, 1997) was administered. The instrument covers all necessary domains of behaviour and cognition in FTLD. The family caregiver and the patient were interviewed separately. The value of the instrument is that it allows the clinician to tailor the global assessment to the particular strengths and weaknesses of the individual patient. This assessment required about 20 min of patient time and 20 min of caregiver time.

Functional assessment

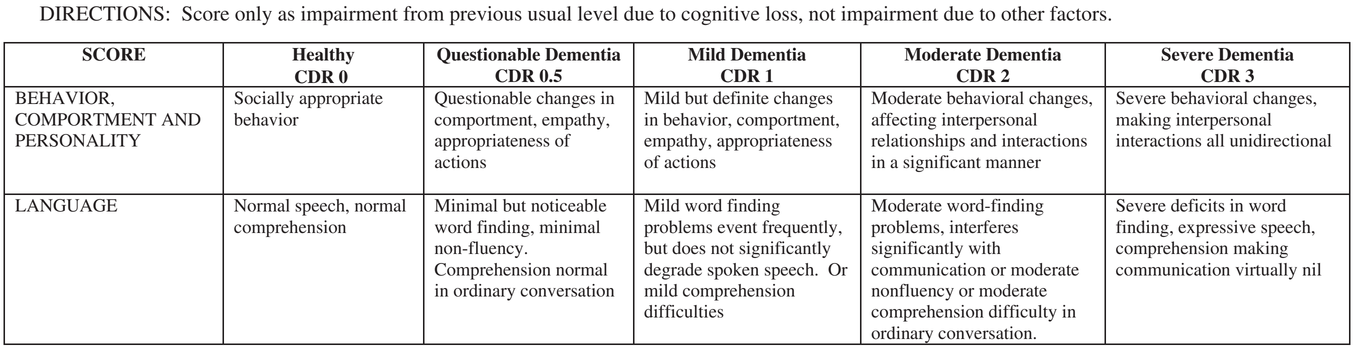
Functional assessments are required in Alzheimer therapeutic trials by regulatory agencies, and are likely to be required for FTLD trials as well. The functional assessment questionnaire (Pfeffer *et al.*, 1982) Functional Activities Questionnaire (FAQ) was chosen because it was a concise instrument that also had been used extensively. In the FAQ, informants rate the patient on 10 complex, higher-order activities, including money management, cooking, shopping, recreation, awareness and memory and ability to use transportation. Higher scores reflect greater degree of impairment.

FTLD-specific Clinical Dementia Rating (FTLD-CDR)

The CDR has been widely used in Alzheimer therapeutic trials. It serves as both a functional assessment instrument and a global rating. Therefore, we administered the standardized CDR according to standard methods and scored according to published methods (Morris, 1993). In addition, we developed two additional domains—Language and Behaviour, Comportment and Personality—in order to capture key FTLD patient characteristics (Fig. 1) that are not explicitly measured by the standard CDR. The two new domains were structured in a parallel fashion to the standard domains and scored on the same scales. These domains were incorporated into a sum of scores of all domains (called the 'sum of Boxes') but not used to generate a 'global' rating. In order to complete the FTLD-CDR, both the family informant and the patient were interviewed using the same semi-structured interview used for the CGIC.

Behavioural scales specific for FTLD

Rating scales for behaviour relevant to the FTLD's were available. The Frontal- Behavioral Inventory (FBI) (Kertesz *et al.*, 1997; Kertesz *et al.*, 2000) and the Neuropsychiatric Inventory (NPI) (Cummings *et al.*, 1994) were administered. The FBI is a 24-item



DIRECTIONS: Score only as impairment from previous usual level due to cognitive loss, not impairment due to other factors.

SCORE	Healthy CDR 0	Questionable Dementia CDR 0.5	Mild Dementia CDR 1	Moderate Dementia CDR 2	Severe Dementia CDR 3
BEHAVIOR, COMPORTMENT AND PERSONALITY	Socially appropriate behavior	Questionable changes in comportment, empathy, appropriateness of actions	Mild but definite changes in behavior, comportment, empathy, appropriateness of actions	Moderate behavioral changes, affecting interpersonal relationships and interactions in a significant manner	Severe behavioral changes, making interpersonal interactions all unidirectional
LANGUAGE	Normal speech, normal comprehension	Minimal but noticeable word finding, minimal non-fluency. Comprehension normal in ordinary conversation	Mild word finding problems event frequently, but does not significantly degrade spoken speech. Or mild comprehension difficulties	Moderate word-finding problems, interferes significantly with communication or moderate nonfluency or moderate comprehension difficulty in ordinary conversation.	Severe deficits in word finding, expressive speech, comprehension making communication virtually nil

Fig. 1 The Behaviour–Comportment–Personality and Language domains of the FTLD–CDR. Ratings range from '0' for normal function, '0.5' for questionable or very mild abnormalities, to '3' indicating severe disturbances.

questionnaire that assesses various behaviours commonly aberrant in FTD that is completed by the caregiver. The NPI is a 12-item instrument also completed by the caregiver. The NPI assesses neuropsychiatric symptoms including depression, anxiety, agitation, euphoria, apathy, disinhibition, irritability, aberrant motor activity, sleep disturbances, hallucinations, delusions and anxiety.

Cognitive instruments

The choice of cognitive instruments was dictated by the availability of information on their performance in the FTLD population (Kramer *et al*., 2003) as well as pragmatic and logistical considerations. We wished to limit the battery to under 1 hour. We wished to have a mix of tasks that required verbal responses and those that were performance-based. There were several features that were required. (i) The tests chosen were not expected to exhibit floor effects at the baseline visit; (ii) the tests should not be so easy that subjects would perform at ceiling at the baseline visit; and (iii) a test such as the Wisconsin Card Sort was eliminated because it is not suited to repeated use.

The cognitive battery included the following tasks:

(i) Learning and recall—the 9-item word delayed recall from California Verbal Learning task (Delis *et al*., 2000). There were four learning trials followed by a free recall and recognition session 30 min later. Because learning and recall are typically preserved in mildly affected FTLD patients (Perry and Hodges, 2000; Rascovsky *et al*., 2002), learning and recall might serve as a good means of measuring longitudinal change.

(ii) Visual confrontation naming—30 items from Boston Naming test (Kaplan *et al*., 1978). Impaired confrontation naming is a principal deficit in many FTLD patients, but the expectation was that it still might be a useful longitudinal measure.

(iii) Verbal fluency—Patients were asked to produce as many words as they could in 1 min for each of three semantic categories and three letters. Fluency tasks are easy to administer and are not time-consuming. As timed tests, they measure speed of performance, as well as aspects of linguistic and executive function.

(iv) Verbal similarities—The verbal similarities subtest from Wechsler Adult Intelligence Scale-revised (WAIS-R) (Wechsler, 1981). The verbal similarities subtest was chosen as a measure of semantic knowledge. The expectation was that some patients would do very poorly at baseline, but enough subjects would perform above floor levels in order for it to be a useful longitudinal measure.

(v) Digits Backward—Items from Wechsler Memory Scale (Wechsler, 1987). Digits backwards is easy to administer and takes little time. Previous studies have found this test useful as a measure of verbal agility (Kramer *et al*., 2003).

(vi) Number cancellation—A 2-number cancellation task drawn from the Alzheimer's Disease Cooperative Study (Mohs *et al*., 1997). The number of correctly crossed out numbers in 60 s was scored. Number cancellation was one of several non-verbal measures of mental agility including the Stroop test and digit symbol substitution that were expected to assess the function longitudinally.

(vii) Stroop Test—The standard Stroop colour and colour-word test. The number of correct colour words named was tabulated.

(viii) Digit symbol substitution—The digit symbol substitution task from WAIS-R (Wechsler, 1981).

(ix) Simplified Trailmaking—A modification of the standard Trailmaking task was devised that involves alternation between numbers and days of the week, and thus requiring only 14 connecting lines. Three outcomes were recorded, the time to complete all items (maximum 120 s), the number of items completed in 120 s, and the number of errors. We expected that the standard Trailmaking test part B (Reitan, 1958) would result in far too many patients unable to complete the task), so we used a modified version that we believed would be easier.

(x) Spontaneous speech—Spontaneous speech rated according to Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). Results of this task will not be reported here, as they proved difficult to standardize across centres.

(xi) MMSE (Folstein *et al*., 1975)—A standard MMSE using 'WORLD' backwards as the concentration item was administered.

Statistical methods

General

Mean values and standard deviations were computed for each assessment instrument by syndrome and for the group as a whole. Although the majority of patients were seen for the follow-up visit in the time frame of 12–13 months after their baseline visit, there were a few outliers who were seen later. We excluded patients from longitudinal analyses whose follow-up exceeded 16 months, and we annualized all scores for those seen between 12 and 16 months. Annualized change scores were computed for all instruments except the CGIC by multiplying the change score by (12/time interval between baseline and follow-up visit in months). The number of patients who scored at the highest level (ceiling) and at the worst level (floor) were tabulated. We did not perform statistical tests on differences between syndromic groups because (i) it was not our focus, and (ii) we did not believe we could equate general severity across groups.

Cognitive composite

In order to reduce the number of cognitive assessment variables, we used a factor analytic approach to validate grouping of the cognitive tests into a small number of domains. A principal components factor analysis with varimax rotations was used. We required a factor loading of 0.5 or greater as a marker of a variable loading on a particular factor. Because the factor weights were likely to be specific for this dataset, the results of the factor analysis were used only to assign tests to a cognitive composite category. We expected that there would be two or at most three factors identified.

Cognitive composites were calculated by averaging the *Z*-scores on the constituent variables defined by the factor analysis and transforming the scores to have a mean of 100 and standard deviation of 15. Each cognitive test included in a composite contributed equally; no weighting of tests was used. Variables that were not normally distributed were log- or square-root transformed to produce a normal distribution. Patients with missing data were excluded from these analyses. We did not attempt to weight certain tests more or less depending upon their longitudinal performance. The *Z*-scores were based on the baseline scores of the current sample and thus should not be interpreted as reflecting normative performance.

Results

Baseline assessments

Of 118 patients who were formally screened for participation, 107 patients completed the baseline assessment, 2 had other diagnoses and 9 were unable to complete the baseline evaluations. Of the 107, 47 had bvFTD, 25 PNFA, 9 PLA and 26 SMD. There was very high agreement between our operationalized criteria and Neary criteria for bvFTD (FTD by Neary nomenclature), with 44 of 47 of our bvFTD cases also meeting Neary criteria. For PNFA (progressive aphasia by Neary nomenclature), there was perfect agreement. Among the 26 patients diagnosed with SMD, 23 met Neary criteria for SMD; the other 3 did not. Of the nine PLA patients, eight met Neary criteria for progressive aphasia. Because of the concern that PLA might be more likely to represent Alzheimer's disease, we have tabulated results with the PLA patients excluded from group data.

The patients who completed the baseline assessment had a mean age of 62.6 ± 9.2 years (median 62.5 years; range 33–83 years). There were 55 men, 52 women. The mean duration of symptoms was 4.3 ± 3.0 years (median 3 years; range 1–20 years). The mean education was 15.0 ± 2.4 years (range 7–20). Baseline MMSE was 23.8 ± 4.9 (range 8–30). The distribution of MMSE scores is shown in Fig. 2A.

Descriptive statistics for the global assessments, functional and behavioural assessments and cognitive instruments are broken down by syndrome (Table 1). For the 'sum of boxes,' the sum of ratings of all domains, on the FTLD-CDR, there was a wide range of severity. No patients were rated as unimpaired ('0') on all domains, and none were rated as severely impaired. The distribution of the FTLD-CDR sum of boxes is shown in Fig. 2B. The FTLD-CDR Behaviour, Comportment and Personality domain and the Language domain appeared to offer additional information. At the very mild end of the spectrum, abnormal ratings for the Language domain were particularly notable. Among 29 bvFTD, SMD or PNFA patients with a standard CDR sum of boxes of ⩽1.5, there were 17 patients with ratings on the Language domain of 1 or greater, including eight with ratings of 2 or greater. On the Behaviour, Comportment and Personality domain in the 29 patients rated on the standard CDR sum of boxes of ⩽1.5, five had ratings of 1 or higher. Ratings of 2 or higher on the Behaviour, Comportment and Personality domain were common among more impaired patients, with standard CDR sum of boxes scores of >2. Eight of the nine PLA patients had Language domain ratings of ⩾1, although none had a standard CDR sum of boxes exceeding 3.5.

For the three rating scales completed by informants—the FAQ, FBI and NPI—there were no ratings indicative of maximal impairment (floor values). A few patients were rated as having no impairment (ceiling values) on the FAQ ($n = 8$), FBI ($n = 2$) and NPI ($n = 12$). There was substantial variability by syndrome on each scale. Consistent with expectations based on the syndromic definitions, the bvFTD and SMD patients had the highest scores on the FAQ, NPI and FBI, while PNFA and PLA were lower on all three. On the NPI, with the exception of hallucinations and delusions (which were quite infrequent), all other 10 items were frequently rated as abnormal.

Data from the baseline performance on cognitive instruments is shown in Table 1. There were differences across syndromes, as expected on language-based tests. There were several instances of floor and ceiling effects on the individual tests such as recognition memory and naming (ceiling values), and simplified Trailmaking (ceiling effects) and delayed recall (floor values) (Table 2). Performance on the individual cognitive tests varied across syndromes in patterns that generally matched expectations. For example,

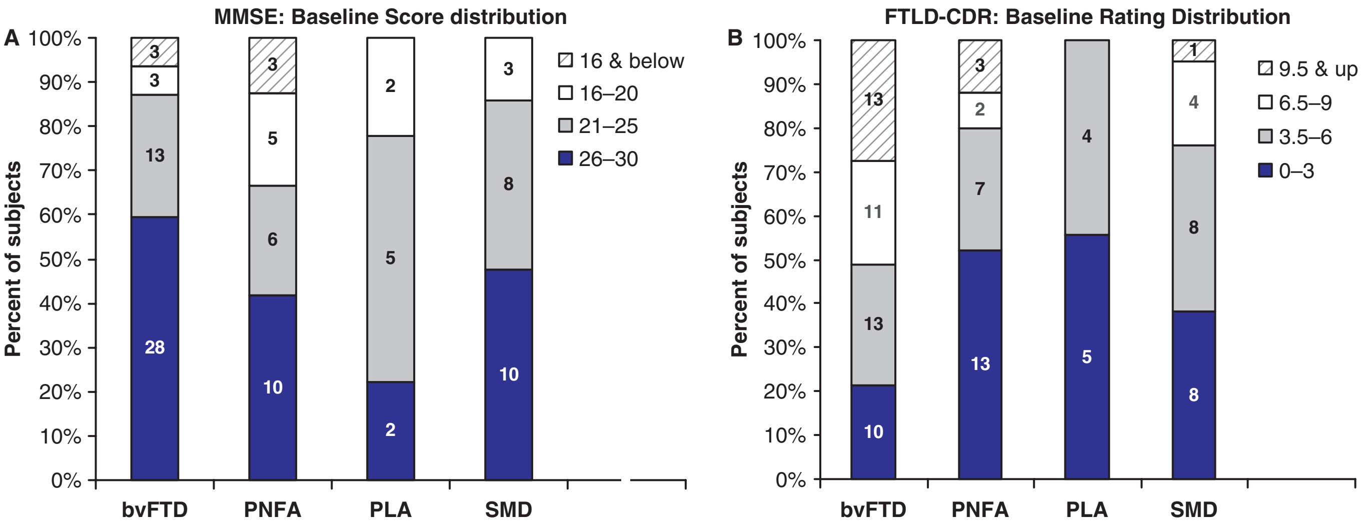


Fig. 2 (**A**) Distribution of baseline MMSE scores by diagnosis. (**B**) Distribution of baseline FTLD–CDR scores by diagnosis. For both histograms, the mildest range of impairment is at the base of each bar, with successively more impaired range above. The actual numbers of subjects are displayed within the bars.

Table 1 Values at baseline assessment on FAQ, FBI, FTLD-CDR, standard CDR, NPI, and cognitive tests by syndrome

Instrument	bvFTD		PNFA		PLA		SMD		ALL[a]	
	N	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)
FTLD-CDR (eight domains)	47	6.79 (3.60)	25	4.18 (2.88)	9	2.94 (1.74)	26	6.00 (4.51)	98	5.91 (3.82)
Standard CDR (six domains)	46	4.64 (3.09)	25	2.26 (2.25)	9	2.83 (2.25)	26	3.85 (3.71)	97	3.81 (3.20)
FAQ	46	16.22 (8.86)	24	7.62 (7.58)	9	10.33 (8.47)	26	11.46 (8.99)	96	12.78 (9.24)
FBI	43	26.81 (12.72)	25	12.00 (9.05)	9	8.44 (5.25)	26	17.69 (12.36)	94	20.35 (13.24)
NPI	46	30.17 (18.8)	25	10.68 (12.48)	9	5.56 (3.91)	26	20.50 (19.25)	97	22.56 (19.15)
Sum of learning over four trials	47	19.72 (6.34)	25	16.56 (10.19)	9	13.44 (8.57)	26	13.38 (7.47)	98	17.23 (8.14)
Delayed free recall	47	3.26 (2.90)	24	4.17 (2.97)	9	2.67 (2.87)	26	1.65 (2.38)	97	3.05 (2.91)
Delayed recognition	47	7.91 (1.23)	24	7.42 (1.98)	9	8.44 (0.73)	26	5.92 (2.42)	97	7.26 (1.97)
Trailmaking (s)	47	66.66 (39.54)	24	80.38 (40.62)	9	102.33 (29.29)	26	56.65 (33.38)	97	67.37 (38.84)
Trailmaking (number correct)	47	11.45 (4.34)	24	9.88 (5.15)	9	10.78 (4.63)	26	12.85 (2.84)	97	11.43 (4.32)
Number cancellation (number correct)	44	26.16 (11.47)	25	21.52 (10.91)	9	22.11 (9.39)	26	28.27 (7.96)	95	25.52 (10.67)
Digits backwards	46	3.87 (1.51)	24	2.46 (1.25)	9	2.78 (1.20)	26	4.31 (1.29)	96	3.64 (1.55)
Stroop colour word (number correct)	43	44.26 (21.97)	23	31.48 (16.82)	9	33.89 (13.83)	26	40.42 (20.82)	92	39.98 (20.91)
Digit symbol substitution	45	45.84 (18.79)	24	38.12 (18.41)	9	38.33 (20.57)	26	51.73 (19.27)	95	45.51 (19.28)
Letter fluency	45	22.98 (13.71)	23	12.83 (8.68)	9	18.78 (9.60)	26	21.00 (11.46)	94	19.95 (12.62)
Category fluency	47	25.32 (10.91)	24	20.42 (11.70)	9	19.44 (9.58)	26	14.35 (12.73)	97	21.16 (12.38)
Confrontation naming	46	23.37 (6.37)	24	19.54 (9.02)	9	14.22 (8.98)	25	6.12 (5.97)	95	17.86 (10.03)
Verbal similarities	44	15.16 (6.33)	23	14.00 (8.17)	9	10.44 (4.56)	26	9.81 (8.00)	93	13.38 (7.57)
MMSE	47	25.13 (4.33)	25	22.04 (6.07)	9	22.33 (3.67)	26	23.73 (4.65)	98	23.97 (5.02)

FTLD-CDR = 8-domain CDR scale; Standard CDR = 6-domain CDR scale.
[a]ALL patients excluding PLA.

Table 2 Neuropsychological test battery instances of floor and ceiling values, number of cases

	N Baseline	*N* Follow-up	Ceiling Baseline (%)	Ceiling Follow-up (%)	Floor Baseline (%)	Floor Follow-up (%)
Sum of learning over four trials	98	68	0.0	0.0	1.0	2.9
Delayed free recall	97	68	4.1	4.4	30.9	52.9
Delayed recognition	97	68	32.0	29.4	0.0	7.4
Trailmaking time to complete	97	67	0.0	0.0	27.8	31.3
Trailmaking (number correct)	97	68	68.0	63.2	1.0	4.4
Number cancellation (number correct)	95	63	3.2	1.6	2.1	4.8
Digits backwards	96	67	2.1	0.0	4.2	16.4
Stroop colour word (number correct)	92	54	0.0	1.9	0.0	3.7
Digit symbol substitution	95	63	0.0	0.0	0.0	6.3
Category or letter fluency	94	65	0.0	0.0	2.1	3.1
Confrontation naming	95	62	9.5	6.5	3.2	4.8
Verbal similarities	93	58	0.0	0.0	6.5	13.8
MMSE	98	68	8.2	2.9	0.0	1.5

All subjects including PLA were included.

SMD and PLA patients had the worst naming, verbal similarities and category fluency performance, while the SMD patients (but not the PLA patients) had the best digit symbol substitution, simplified Trailmaking and letter cancellation scores. PNFA patients had the lowest performance on letter fluency of the four syndromic groups. bvFTD patients had scores that were intermediate relative to the other syndromes.

Longitudinal outcomes

There were 90 (84%) who completed the follow-up assessments. Seventeen patients who had been fully evaluated at baseline were lost to follow-up because of death ($n = 1$) or unwillingness to continue in the study ($n = 16$). Of those who were evaluated longitudinally, one was evaluated at 10.1 months, 77 were evaluated between 11 and 16 months after the initial visit and 12 were seen >16 months after the initial visit. Those seen >16 months post-baseline were excluded from the longitudinal analyses. The patients who were part of the longitudinal analyses included 78 patients (36 with bvFTD, 16 with PNFA, 9 with PLA and 17 with SMD). The mean time between baseline and follow-up for patients in the longitudinal analysis was 12.5 (±1.1) months (range 10.1–15.9 months). There were no differences on any baseline measure between the

Table 3 Values for annualized longitudinal changes for FAQ, FBI, FTLD-CDR, standard CDR, CGIC, NPI and cognitive tests

Instrument	bvFTD		PNFA		PLA		SMD		ALL[a]	
	N	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)	*N*	Mean (SD)
FTLD-CDR (eight domains)	36	3.4 (3.75)	15	3.53 (3.91)	9	4.39 (5.21)	17	3.57 (2.46)	68	3.47 (3.46)
Standard CDR (six domains)	36	2.71 (2.81)	15	2.74 (3.52)	9	3.44 (4.33)	17	2.77 (1.88)	68	2.73 (2.75)
Clinical global impression	30	5.27 (0.88)	13	5.28 (0.86)	9	6.01 (0.59)	13	4.95 (1.36)	56	5.2 (1.00)
FAQ	35	5.81 (6.54)	16	5.76 (5.32)	9	6.31 (5.92)	15	7.61 (4.95)	66	6.21 (5.9)
FBI	34	6.54 (10.57)	16	4.95 (7.98)	9	6.19 (6)	15	7.41 (9.41)	65	6.35 (9.63)
NPI	35	5.84 (18.53)	16	4.13 (6.76)	9	6.11 (9.45)	15	10.13 (10.28)	66	6.4 (14.75)
Sum of learning over four trials	34	−2.63 (4.60)	12	−1.85 (2.94)	9	−2.57 (4.05)	17	−1.56 (3.00)	63	−2.2 (3.91)
Delayed free recall	34	−0.5 (2.17)	12	−0.48 (0.98)	9	−1.48 (2.30)	17	−0.37 (0.93)	63	−0.46 (1.70)
Delayed recognition	34	−0.47 (1.38)	12	0.24 (1.39)	9	−1.34 (1.89)	17	−0.83 (2.31)	63	−0.43 (1.69)
Trailmaking (s)	33	10.77 (28.8)	12	11.66 (28.75)	9	10.38 (12.23)	17	7.75 (41.97)	62	10.11 (32.38)
Trailmaking (number correct)	34	−1.17 (3.45)	12	0.72 (2.62)	9	−5.03 (4.91)	17	−0.89 (4.57)	63	−0.74 (3.67)
Number cancellation (number correct)	31	−2.75 (8.81)	11	−4.54 (4.40)	9	−4.62 (4.44)	17	−5.51 (8.34)	59	−3.88 (8.01)
Digits backwards	34	−0.69 (1.13)	13	−0.46 (1.10)	9	−0.35 (0.70)	17	−0.59 (1.32)	64	−0.62 (1.16)
Stroop colour word (number correct)	29	−7.53 (16.89)	11	−15.26 (12.79)	9	−13.35 (9.59)	17	−15.68 (12.85)	57	−11.45 (15.34)
Digit symbol substitution	32	−6.54 (10.21)	11	−2.63 (10.78)	9	−12.65 (15.92)	17	−5.33 (16.11)	60	−5.48 (12.13)
Letter fluency	34	−6.21 (7.80)	13	−2.48 (5.00)	9	−6.41 (7.19)	17	−5.87 (4.71)	64	−5.36 (6.67)
Category fluency	34	−6.4 (5.95)	12	−7.92 (6.99)	9	−6.91 (5.16)	17	−3.53 (4.97)	63	−5.92 (6.03)
Confrontation naming	32	−2.07 (4.12)	11	−3.16 (1.84)	9	−3.49 (4.26)	17	−1.15 (2.43)	60	−2.01 (3.40)
Verbal similarities	31	−2.91 (5.75)	10	−5.02 (9.29)	9	−2.36 (3.30)	17	−3.08 (3.21)	58	−3.32 (5.88)
MMSE	33	−2.45 (4.32)	14	−4.95 (6.12)	9	−6.37 (4.24)	16	−7.53 (4.43)	63	−4.29 (5.18)

[a]ALL patients excluding PLA.

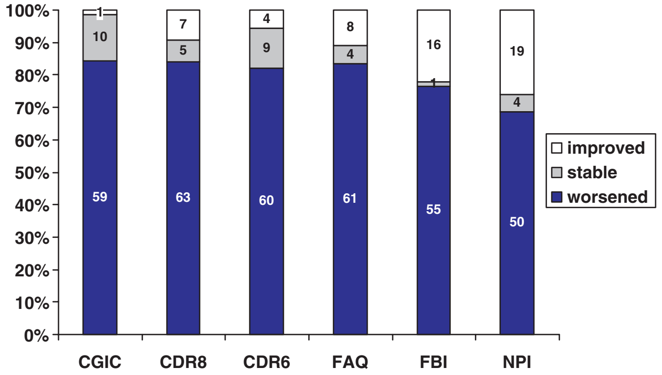


Fig. 3 Longitudinal changes in Rating Scales. Proportion of bvFTD, PNFA and SMD patients who were worse, stable or improved on the global (CGIC, FTLD–CDR and standard CDR) and functional measures (FAQ, FBI, NPI) over the course of the 12–16 month follow-up period.

27 subjects who were not included in the longitudinal analyses and those who were.

Descriptive statistics for the annualized longitudinal changes on global assessments, functional and behavioural assessments and cognitive instruments were broken down by syndrome (Table 3, Fig. 3). The CGIC ratings showed that 59 (84%) bvFTD, SMD or PNFA patients worsened over 12 months. Only one patient (with bvFTD) was rated as improved and 10 (14%) were rated as unchanged. All of the PLA patients were rated as worse on the CGIC. The distribution of ratings was similar across syndromes. The FTLD-CDR 'sum of boxes' ratings similarly showed worsening in the overwhelming majority of patients. The average increase in the rating of dementia severity was −3.6 ± 3.7 rating points. The change scores on the FTLD-CDR sum of boxes exceeded that of the standard CDR sum of boxes in 40 of 67 (60%) of bvFTD, SMD or PNFA patients and in six of nine PLA patients. On the FTLD-CDR, 63 (84%) bvFTD, SMD or PNFA patients were rated as worse, 5 (7%) as unchanged and 7 (9%) patients were rated as less impaired than at baseline. Six of those rated as improved had bvFTD, and one had PNFA as syndromic diagnoses.

The FAQ showed a similar proportion of patients who worsened, remained the same or improved. In contrast the FBI and NPI both had a somewhat higher proportion of patients who were rated as improved at 1 year. The mean declines on the FAQ were similar for the four syndromes. On the NPI and the FBI, the SMD patients showed more worsening than the other three groups.

The annualized changes in cognitive test scores across syndromes are shown in Table 3. The number of instances of floor and ceiling values for testing at month 12 is shown in Table 2. On the MMSE, the average decline across all bvFTD, SMD and PNFA patients was 4.3 ± 5.2. The PLA patients exhibited greater decline (6.4 ± 4.2) on the MMSE. A number of patients were unable to complete at least one of the tests (in particular, the simplified Trailmaking, delayed free recall and verbal similarities), although fewer patients still performed at ceiling levels on other tests (naming, delayed free recall). Numerically, all syndromic groups showed decline on all instruments, and no one syndromic group appeared to worsen on more tests than any other.

Table 4 Values of cognitive composites at baseline, month 12 and difference scores, by syndrome

	bvFTD mean (SD)	PNFA mean (SD)	PLA mean (SD)	SMD mean (SD)	ALL[a] mean (SD)
N	27	9	9	17	53
Executive composite baseline	105.77 (15.44)	96.97 (17.04)	92.51 (13.18)	102.49 (8.98)	103.22 (14.11)
Executive composite 12 months	100.02 (17.57)	91.29 (18.23)	82.43 (14.19)	95.96 (12.60)	97.23 (16.28)
Executive composite difference	−5.75 (7.37)	−5.68 (6.37)	−10.08 (6.06)	−6.53 (9.28)	−5.99 (7.75)
Language composite baseline	110.65 (12.46)	105.23 (16.17)	95.91 (12.73)	89.53 (13.22)	102.96 (16.19)
Language composite 12 months	104.29 (15.42)	99.04 (16.39)	88.57 (11.72)	83.69 (12.18)	96.79 (17.09)
Language composite difference	−6.36 (6.35)	−6.20 (3.65)	−7.34 (5.52)	−5.85 (4.49)	−6.17 (5.33)
Global composite baseline	108.21 (12.72)	101.10 (16.01)	94.21 (11.86)	96.01 (9.54)	103.09 (13.38)
Global composite 12 months	102.16 (15.45)	95.17 (16.87)	85.50 (12.15)	89.82 (11.13)	97.01 (15.25)
Global composite difference	−6.05 (5.92)	−5.94 (2.75)	−8.71 (4.96)	−6.19 (4.68)	−6.08 (5.04)

Only patients with complete data for all measures were included.
[a]ALL patients excluding PLA.

We calculated differences between groups on all tests in Table 3 using an ANOVA test. Only two assessments, the number correct on simplified Trailmaking ($P=0.02$) and the MMSE ($P<0.001$) showed group differences, but only the latter survived a Bonferroni correction. Descriptively, the group difference in the simplified Trailmaking was accounted for by the stability of performance in the PNFA group, whereas all other groups declined. The group difference on the MMSE was accounted for by the large declines in the SMD and PLA groups and the lesser change in the bvFTD group.

Cognitive composite

Eleven variables were selected as candidate variables for inclusion in the factor analysis: immediate recall (sum of learning over four trials from California Verbal Learning test), simplified Trailmaking (ratio of number of correct lines per minute divided by time to complete), Boston Naming test number correct, backward digit span length, Digit Symbol Substitution task score, total letter fluency, total category fluency, number correct on Stroop interference test, time to complete number cancellation, verbal similarities correct and delayed word list recognition. In addition, a sum-of-errors variable was constructed from three tests (Stroop interference test, simplified Trailmaking and delayed recognition). Three variables—Stroop, simplified Trailmaking ratio and the error composite—required square root or log transformations to produce a normal distribution. The data from the baseline assessments for these 11 variables were analyzed with a principal components factor analysis with varimax rotations.

The factor analysis yielded two factors, one with six variables (simplified Trailmaking ratio, backward digit span, Digit Symbol, Stroop, number cancellation and sum-of errors (described above) comprising a working memory/ executive factor and another with five variables (Boston Naming test, the two verbal fluency scores, Similarities and sum of learning over four trials from California Verbal Learning test) comprising a verbal factor. We repeated the principal components analysis 10 times, each time leaving out a different 10% of the sample. Each of the 10 cross-validation analyses yielded the same pattern of variables loading on the two factors, supporting the robustness of the solution and our decision to select those variables for construction of the composite scores. The factor analysis provided the basis for grouping the variables into two composites. We refer to the composite created from the first factor as the executive composite and the composite created from the second factor as the language composite. As described in methods, the composites were created from the *Z*-scores from the baseline values of each of the 11 variables.

The values of the composite scores—the executive composite, the language composite and the global composite across syndromes at baseline are shown in Table 4, as are the annualized change scores. At baseline, the bvFTD and SMD patients had higher scores on the executive composite than the other two groups. The bvFTD group was also the highest on the language composite, but the SMD group was the lowest. The individual test scores (Table 1) show the basis for these patterns. For example, the sum of learning trials was nearly 50% higher in the bvFTD group compared to the SMD group. The PNFA group was substantially impaired on letter fluency compared to either bvFTD or SMD patients.

Over the group as a whole, they declined 6.1 units on the combined composite (Table 4), which was equivalent to about half of a standard deviation of the baseline composite (Cohen $d=0.5$). Patients diagnosed with PLA declined the most, followed by bvFTD, while those with PNFA declined the least. The data on individual tests in Table 3 reflect the pattern of decline seen across syndromes.

The changes on the combined cognitive composite were correlated with the change on the CDR (Pearson $r=0.40$, $P=0.001$), change on the FAQ (Pearson $r=-0.20$, $P=0.128$) and MMSE (Pearson $r=0.26$, $P=0.069$).

Power considerations

Table 5 shows the numbers of patients needed in a randomized, parallel group, placebo-controlled clinical trial

Table 5 Power estimates for FTLD-CDR (Non-annualized) and global cognitive composite for small and conservative medium sized effects—sample sizes per group, change scores (β = Power, α = 0.05, two sample)

	Change in FTLD-CDR sum of boxes				Change in global composite score			
	Mean (SD)	Effect size of mean	$\beta = 0.80$	$\beta = 0.90$	Mean (SD)	Effect size of mean	$\beta = 0.80$	$\beta = 0.90$
bvFTD	3.40 (3.75)	Small	307	410	−6.05 (5.92)	Small	242	323
		Medium	121	161		Medium	95	127
PFNA	3.53 (3.91)	Small	310	414	−5.94 (2.75)	Small	55	74
		Medium	122	163		Medium	23	30
PLA	4.39 (5.21)	Small	355	475	−8.71 (4.96)	Small	83	111
		Medium	140	186		Medium	33	44
SMD	3.57 (2.46)	Small	121	161	−6.19 (4.68)	Small	145	194
		Medium	48	64		Medium	58	77
ALL[b]	3.47 (3.46)	Small	251	336	−6.08 (5.04)	Small	174	233
		Medium	99	132		Medium	69	92

[a]Effect size of mean = small (25%), conservative medium (40%).
[b]ALL patients excluding PLA.

to demonstrate efficacy on the global cognitive composite and the FTLD-CDR, assuming small (25%) or conservative moderate (40%) effect sizes. For studies involving bvFTD, PNFA and SMD, over 100 subjects per group in a 2-arm placebo controlled trial would be needed to achieve at least 80% power to detect a moderate sized effects for both outcome measures.

Discussion

This study has a number of key findings. First, we showed that it was feasible to recruit and follow patients with FTLD in a multicentre study. Second, we showed that functional, global and cognitive instruments could be administered to FTLD patients in a longitudinal fashion that replicated the practices of a clinical trial. We further demonstrated that some instruments worked well, others not as well, and that there is a basis for designating primary and secondary outcome measures in future clinical trials in FTLD. Fourth, we showed that the instruments performed qualitatively similar though quantitatively different across the syndromes of FTLD. Finally, our study provides quantitative information about change and variability that can be used by other investigators for designing future clinical trials.

We succeeded in recruiting 107 patients with mild to moderate impairment from five centres over a 2-year period. One centre found that they did not have the referral base that they had anticipated to identify the target number of patients. Nearly 80% of patients returned for the follow-up visit 12–16 months after the baseline visit, despite the fact that there was no therapeutic benefit for them. As a group, our patients showed a moderate amount of decline over the 12 months of observation.

Because the current consensus criteria for FTLD (Neary *et al.*, 1998) were not intended for clinical trials, we developed explicit inclusion and exclusion criteria based on the consensus criteria, but also taking newer observations into account. In addition to the clinical criteria, all patients in this study were required to have imaging studies that showed focal atrophy in structures implicated in FTLD. These operational criteria for the syndromes were shown to have concurrent validity with published criteria. In fact, there was nearly perfect correspondence between our operational criteria for PNFA, SMD and bvFTD with the consensus criteria. The pattern of cognitive and behavioural changes that distinguished bvFTD, PNFA, PLA and SMD in our study were very similar to observations of others (Hodges *et al.*, 1999; Kramer *et al.*, 2003; Gorno-Tempini *et al.*, 2004; Libon *et al.*, 2007). However, we do not have autopsy confirmation on anyone in this study, as of May 2008. Furthermore, our patients almost uniformly showed decline over time, suggesting that our operationalized entry criteria successfully identified a group with progressive disease.

A major goal of this project was to assess various functional, cognitive and global instruments for use in clinical trials in the FTLD cognitive disorders. We chose instruments that had face validity for FTLD in cross-sectional investigations of FTLD and that appeared to be suitable for longitudinal purposes. There were many differences in rate and pattern of decline across the FTLD clinical syndromes. PNFA patients generally showed the least decline, while patients with the other syndromes showed greater declines. Patients with SMD demonstrated the most rapid declines, although the PLA patients had nearly the same degree of decline. While the differences in functional assessment and cognitive test scores across syndromic groups were not statistically significant due to the relatively small sample sizes and intrinsic between-subjects variability, the data suggest that it is feasible to use a common cognitive battery across the different syndromes.

Some instruments were specifically developed for this project. This is the first study, to our knowledge, that used global instruments such as the CGIC and a modified CDR to assess change in FTLD disorders. We showed that both a CGIC and an FTLD-modified-CDR had suitable properties for a clinical trial. The average decline on the FTLD-CDR was 3.47 points. The additional domains of the FTLD-CDR added unique information, especially in patients with very mild impairment. The Language domain rating was particularly useful in many patients who were rated as having a standard CDR sum of boxes of 0. The Behaviour–Comportment–Personality Domain tracked closely with the standard CDR sum of boxes, but provided unique information about severity in patients who otherwise would have been rated as only mildly impaired. Using only the standard six domains of the CDR, the average decline was 2.73. These values are slightly larger than were seen in the placebo group of a 1 year trial of NSAIDs in Alzheimer's disease (Aisen *et al.*, 2003), where the average decline on the 6-item standard CDR was 2.2. In a 6-month trial of donepezil in Alzheimer's disease, the mean decline on the 6-item CDR in the placebo group was only 0.58 (Rogers *et al.*, 1998). These comparisons should be viewed with caution because baseline severity between the Alzheimer's disease patients in the NSAID trial and the current group of FTLD patients were not and could not be equated. The baseline MMSE scores in FTLD patients were higher (23.8) compared to the Alzheimer's disease patients in either the NSAID trial (20.9) (Aisen *et al.*, 2003) or the donepezil trial (20.9) (Rogers *et al.*, 1998).

The longitudinal performance of the functional instruments was less than ideal for use in clinical trials. The FBI and NPI, instruments that assess behavioural abnormalities, yielded expected scores at baseline that differentiated the syndromes. However, a number of our FTLD patients exhibited less disruptive behaviours over time as their disease progressed. The fact that a number of patients showed improved scores on the FBI or NPI detracts from their suitability as a primary outcome measure in a therapeutic trial. We examined individual items on both the FBI and NPI and could not identify any one set of questions where the better scores were likely to appear. We believe that the 'better' scores seen over time on the FBI and NPI reflect increasing apathy and inertia in our patients, which served to blunt the impact of their disruptive behaviours.

Some of the cognitive tests performed poorly in some syndromes. The trailmaking task suffered from excessive numbers of subjects performing both at ceiling and at floor levels. Delayed recognition was generally too easy, resulting in many subjects performing at ceiling. There was relatively little change in the MMSE in bvFTD patients, which should not be surprising given its lack of focus on executive and attentional tasks. In contrast, the rate of change on the MMSE in the SMD patients might make it useful for clinical trials focussed on that set of patients. Delayed recall seemed rather insensitive to change across the group as a whole, and probably should not be part of the cognitive battery in FTLD clinical trials.

In contrast, several neuropsychological tests showed promise. The sum of learning trials, number cancellation, digit symbol, Stroop colour-word naming and both word fluency tasks showed relatively few floor or ceiling values. Even on these tests, however, some patients showed improvement, which contributed to increased variance in performance. The fluency tasks showed the largest ratios of mean change to standard deviation, suggesting that, among individual tests, measures of fluency might be the most efficient for detecting change in a clinical trial. One option for future clinical trials is to select one or two tests and use them as primary cognitive assessment instruments. The drawback of using only a single test, fluency for example, is that it may capture only a fraction of the cognitive decline over the course of a trial. If a larger set of instruments were to be used, some form of data reduction would be necessary to avoid the problem of multiple outcome measures. Hence, we felt that composite measures were required.

We developed two composite cognitive batteries. Two factor loadings were identified with factor analysis, one that corresponded to the executive domain and another that corresponded to the language domain. The goal of defining a composite cognitive assessment for FTLD was to incorporate information from a variety of cognitive tests, and to combine that information into one cognitive score. A cognitive battery that could be reduced to a single score would then be a suitable primary cognitive outcome measure for a clinical trial. In addition, the use of composite measures would also manage the problem of floor and ceiling values on individual tests, and create a variable with a normal distribution. We chose to use *Z*-scores for each element of the composites without weighting some tests higher and some lower. We considered alternative strategies for transforming individual test scores into an ordinal rating scheme to allow this battery to be adapted by others, but further normative work is needed. Power calculations (Table 5) using the global cognitive composite yielded projected sample sizes that were comparable, though slightly more favourable, than the FTDL-CDR.

The cross-sectional and longitudinal data collected in this study are unique. There are very few other instances in which a large number of FTLD cognitive syndrome patients were administered the identical battery, and none, to our knowledge, that were collected prospectively at multiple sites. A recent report (Kipps *et al.*, 2008) using the Addenbrooke's Cognitive Examination (maximum score = 100) found that those bvFTD patients with focal atrophy on MR scans lost 13.4 (± 13.8) points per year (starting from a score of 73.6) while those with normal MR scans lost only 2.8 points (starting from a score of 89.0). These authors noted that selecting patients with evidence of focal cortical atrophy (as was done in the current study) minimized the inclusion of patients with very slowly progressive or non-progressive disease. Their estimates of power were very similar to ours. An autopsy-based series of patients with confirmed FTLD found that the annual rate of decline on the MMSE was 6.7 points (Rascovsky *et al.*, 2002). The series included a variety of clinical syndromes.

Autopsy series have the advantage of diagnostic certainty, but are biased towards more rapidly progressing cases. In contrast, we found a slightly lower annual decline on the MMSE of 4.3 points. We also found substantial differences across FTLD clinical syndromes. A study of bvFTD and progressive aphasia patients (all subtypes) also reported estimates of longitudinal change in several cognitive tests including fluency, naming and MMSE (Wicklund *et al.*, 2007). There were quantitative differences, as would be expected based on differences in diagnostic definitions and severity of patients between their study and ours. A study of longitudinal cognitive changes in autopsy-proved tau-positive versus tau-negative FTLD cases showed worse performance among tau-negative patients on confrontation naming and recognition memory, and worse performance among tau-positive patients on digits backwards and visual constructions (Grossman *et al.*, 2008).

One of the major challenges facing the FTLD field is to target therapies to specific pathological subtypes of the FTLDs. The present study did not address that issue because we have no clinical markers for specific proteinopathies other than the rough guides that syndrome types provide. Although there may be some features that distinguish tauopathies from TPD-43 proteinopathies in group-wise comparisons (Hu *et al.*, 2007), the bvFTD group cannot be differentiated on clinical grounds according to predominant underlying pathology (Hodges *et al.*, 2004; Forman *et al.*, 2006). In contrast, the PNFA group is likely to be mainly tauopathic disorders (Hodges *et al.*, 2004; Forman *et al.*, 2006; Josephs *et al.*, 2006), and SMD may be more likely to be linked to the TDP-43 or progranulin-mutation-mediated spectrum. Unfortunately, the latter two disorders are not as common as bvFTD, and hence trials that were limited to one or the other of these disorders would be a daunting challenge for recruiting. On the other hand, the syndrome of PLA may represent a mixture of cases with Alzheimer pathology and those with one of the FTLD disorders (Galton *et al.*, 2000; Gorno-Tempini *et al.*, 2004; Kertesz *et al.*, 2005; Knibb *et al.*, 2006; Josephs *et al.*, 2008). For this reason, PLA subjects should be excluded from FTLD trials. Hence, we calculated group means after excluding the PLA patients.

The strengths of this study include its prospective, multicentre design that used standard, operational definitions of the FTLD syndromes. Its weaknesses include the modest number of patients among syndromic subtypes, and the loss to follow-up of over 20%. Our estimates of change are constrained by the limited number of subjects followed longitudinally. We cannot determine whether the change scores on the various cognitive and behavioural measures are equivalent across the range of severity. Nonlinearity is certainly a possibility. In addition, as our study was only of 1 year's duration, our data cannot be extrapolated to longer duration trials. The score distributions used to construct the *Z*-scores for the composites may not be replicable in other samples, especially if different inclusion and exclusion criteria are used. Pathological confirmation was not available in any cases at the time of this report's submission. Further work is needed to optimize the cognitive composites. Scoring procedures that do not depend upon the particular study patients need to be developed.

The ultimate value of this work will only be realized when there are therapeutic agents that are worthy of testing in the FTLD disorders. Even then, instruments and procedures such as proposed here will prove their mettle only through use and success in actual clinical trials.

Acknowledgments

The authors would like to thank the study coordinators at all sites, and all of the patients and their devoted family members for participating in this study. The authors would also like to acknowledge the efforts of Walter Kukull PhD and Duane Beekley, at the National Alzheimer Coordinating Center, University of Washington. Statistical analyses were performed by N.M. under the supervision of D.K.

Funding

National Institute of Aging studies (R01-AG023195 and P50-AG 16574 to Mayo Alzheimer's Disease Research Centre, P30-AG19610 to Arizona ADC, P50-AG016570 to UCLA ADRC, P50-AG023501 to University of California ADRC, U01-AG016976 to National Alzheimer Coordinating Centre, University of Washington).

References

Aisen PS, Schafer KA, Grundman M, Pfeiffer E, Sano M, Davis KL, et al. Effects of rofecoxib or naproxen vs placebo on Alzheimer disease progression: a randomized controlled trial. JAMA 2003; 289: 2819–26.

Boxer AL, Boeve BF. Frontotemporal Dementia Treatment: Current Symptomatic Therapies and Implications of Recent Genetic, Biochemical, and Neuroimaging Studies. Alzheimer Dis Assoc Disord 2007; 21: S79–87.

Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. Neurology 1994; 44: 2308–14.

Deakin JB, Rahman S, Nestor PJ, Hodges JR, Sahakian BJ. Paroxetine does not improve symptoms and impairs cognition in frontotemporal dementia: a double-blind randomized controlled trial. Psychopharmacology (Berl) 2004; 172: 400–8.

Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test. 2nd edn. San Antonio, TX: The Psychological Crop; 2000.

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.

Forman MS, Farmer J, Johnson JK, Clark CM, Arnold SE, Coslett HB, et al. Frontotemporal dementia: Clinicopathological correlations. Ann Neurol 2006; 59: 952–62.

Galton CJ, Patterson K, Xuereb JH, Hodges JR. Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. Brain 2000; 123: 484–98.

Goodglass H, Kaplan E. The assessment of aphasia and related disorders. 2nd edn. Philadelphia: Lea & Febiger; 1983.

Gorno-Tempini ML, Dronkers NF, Rankin KP, Ogar JM, Phengrasamy L, Rosen HJ, et al. Cognition and anatomy in three variants of primary progressive aphasia. Ann Neurol 2004; 55: 335–46.

Grossman M, Xie SX, Libon DJ, Wang X, Massimo L, Moore P, et al. Longitudinal decline in autopsy-defined frontotemporal lobar degeneration. Neurology 2008; 70: 2036–45.

Hodges JR, Davies RR, Xuereb JH, Casey B, Broe M, Bak TH, et al. Clinicopathological correlates in frontotemporal dementia. Ann Neurol 2004; 56: 399–406.

Hodges JR, Patterson K, Ward R, Garrard P, Bak T, Perry R, et al. The differentiation of semantic dementia and frontal lobe dementia (temporal and frontal variants of frontotemporal dementia) from early Alzheimer's disease: a comparative neuropsychological study. Neuropsychology 1999; 13: 31–40.

Hu WT, Mandrekar JN, Parisi JE, Knopman DS, Boeve BF, Petersen RC, et al. Clinical features of pathologic subtypes of behavioral—variant frontotemporal dementia. Arch Neurol 2007; 64: 1611–6.

Josephs KA, Duffy JR, Strand EA, Whitwell JL, Layton KF, Parisi JE, et al. Clinicopathological and imaging correlates of progressive aphasia and apraxia of speech. Brain 2006; 129: 1385–98.

Josephs KA, Whitwell JL, Duffy JR, Vanvoorst WA, Strand EA, Hu WT, et al. Progressive aphasia secondary to Alzheimer disease vs FTLD pathology. Neurology 2008; 70: 25–34.

Kaplan E, Goodglass H, Weintraub S. The Boston Naming Test. 2nd edn. Boston: Lea & Fabiger; 1978.

Kertesz A, Davidson W, Fox H. Frontal behavioral inventory: diagnostic criteria for frontal lobe dementia. Can J Neurol Sci 1997; 24: 29–36.

Kertesz A, McMonagle P, Blair M, Davidson W, Munoz DG. The evolution and pathology of frontotemporal dementia. Brain 2005; 128: 1996–2005.

Kertesz A, Nadkarni N, Davidson W, Thomas AW. The Frontal Behavioral Inventory in the differential diagnosis of frontotemporal dementia. J Int Neuropsychol Soc 2000; 6: 460–8.

Kipps CM, Nestor PJ, Dawson CE, Mitchell J, Hodges JR. Measuring progression in frontotemporal dementia: Implications for therapeutic interventions. Neurology 2008; 70: 2046–52.

Knibb JA, Xuereb JH, Patterson K, Hodges JR. Clinical and pathological characterization of progressive aphasia. Ann Neurol 2006; 59: 156–65.

Knopman D, Boeve B, Caselli R, Graff-Radford NR, Kramer JH, Mendez MF, et al. Longitudinal Tracking of FTLD. Toward developing Clinical Trial Methodology. Alzheimer Dis Assoc Disord 2007; 21: S58–63.

Kramer JH, Jurik J, Sha SJ, Rankin KP, Rosen HJ, Johnson JK, et al. Distinctive neuropsychological patterns in frontotemporal dementia, semantic dementia, and Alzheimer disease. Cogn Behav Neurol 2003; 16: 211–8.

Libon DJ, Xie SX, Moore P, Farmer J, Antani S, McCawley G, et al. Patterns of neuropsychological impairment in frontotemporal dementia. Neurology 2007; 68: 369–75.

Mesulam M, Wicklund A, Johnson N, Rogalski E, Leger GC, Rademaker A, et al. Alzheimer and frontotemporal pathology in subsets of primary progressive aphasia. Ann Neurol 2008; 63: 709–17.

Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord 1997; 11 (Suppl 2): S13–21.

Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology 1993; 43: 2412–4.

Neary D, Snowden JS, Gustafson L, Passant U, Stuss D, Black S, et al. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology 1998; 51: 1546–54.

Perry RJ, Hodges JR. Differentiating frontal and temporal variant frontotemporal dementia from Alzheimer's disease. Neurology 2000; 54: 2277–84.

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: 323–9.

Rascovsky K, Salmon DP, Ho GJ, Galasko D, Peavy GM, Hansen LA, et al. Cognitive profiles differ in autopsy-confirmed frontotemporal dementia and AD. Neurology 2002; 58: 1801–8.

Rascovsky K, Salmon DP, Lipton AM, Leverenz JB, DeCarli C, Jagust WJ, et al. Rate of progression differs in frontotemporal dementia and Alzheimer disease. Neurology 2005; 65: 397–403.

Reitan R. Validity of the Trail-making test as an indication of organic brain damage. Percept Motor Skills 1958; 8: 271–6.

Rogers SL, Farlow MR, Doody RS, Mohs R, Friedhoff LT. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Neurology 1998; 50: 136–45.

Schneider LS, Olin JT, Doody RS, Clark CM, Morris JC, Reisberg B, et al. Validity and reliability of the Alzheimer's Disease Cooperative Study-Clinical Global Impression of Change. The Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord 1997; 11 (Suppl 2): S22–32.

Wechsler D. Wechsler Adult Intelligence Scale-Revised. New York: Psychological Corporation 1981.

Wechsler DA. Wechsler Memory Scale-Revised. New York: Psychological Corporation 1987.

Wicklund AH, Rademaker A, Johnson N, Weitner BB, Weintraub S. Rate of Cognitive Change Measured by Neuropsychologic Test Performance in 3 Distinct Dementia Syndromes. Alzheimer Dis Assoc Disord 2007; 21: S70–8.