## Title

Longitudinal multimodal imaging and clinical endpoints for frontotemporal dementia clinical trials



## Journal

Brain, 142(2)

## ISSN

0006-8950

## Authors

Staffaroni, Adam M
Ljubenkov, Peter A
Kornak, John
et al.

## Publication Date

2019-02-01

## DOI

10.1093/brain/awy319

Peer reviewed

doi:10.1093/brain/awy319 BRAIN 2019: 142; 443–459 | 443

BRAIN
A JOURNAL OF NEUROLOGY

# Longitudinal multimodal imaging and clinical endpoints for frontotemporal dementia clinical trials

**Adam M. Staffaroni,[1] Peter A. Ljubenkov,[1] John Kornak,[2] Yann Cobigo,[1] Samir Datta,[1] Gabe Marx,[1] Samantha M. Walters,[1] Kevin Chiang,[1] Nick Olney,[1] Fanny M. Elahi,[1] David S. Knopman,[3] Bradford C. Dickerson,[4] Bradley F. Boeve,[3] Maria Luisa Gorno-Tempini,[1] Salvatore Spina,[1] Lea T. Grinberg,[1,5,6] William W. Seeley,[1,5] Bruce L. Miller,[1] Joel H. Kramer,[1] Adam L. Boxer[1] and Howard J. Rosen[1]**

Frontotemporal dementia refers to a group of progressive neurodegenerative syndromes usually caused by the accumulation of pathological tau or TDP-43 proteins. The effects of these proteins in the brain are complex, and each can present with several different clinical syndromes. Clinical efficacy trials of drugs targeting these proteins must use endpoints that are meaningful to all participants despite the variability in symptoms across patients. There are many candidate clinical measures, including neuropsychological scores and functional measures. Brain imaging is another potentially attractive outcome that can be precisely quantified and provides evidence of disease modification. Most imaging studies in frontotemporal dementia have been cross-sectional, and few have compared longitudinal changes in cortical volume with changes in other measures such as perfusion and white matter integrity. The current study characterized longitudinal changes in 161 patients with three frontotemporal dementia syndromes: behavioural variant frontotemporal dementia ($n = 77$) and the semantic ($n = 45$) and non-fluent ($n = 39$) variants of primary progressive aphasia. Visits included comprehensive neuropsychological and functional assessment, structural MRI (3 T), diffusion tensor imaging, and arterial spin labelled perfusion imaging. The goal was to identify measures that are appropriate as clinical trial outcomes for each group, as well as those that might be appropriate for trials that would include more than one of these groups. Linear mixed effects models were used to estimate changes in each measure, and to examine the correlation between imaging and clinical changes. Sample sizes were estimated based on the observed effects for theoretical clinical trials using bootstrapping techniques to provide 95% confidence intervals for these estimates. Declines in functional and neuropsychological measures, as well as frontal and temporal cortical volumes and white matter microstructure were detected in all groups. Imaging changes were statistically significantly correlated with, and explained a substantial portion of variance in, the change in most clinical measures. Perfusion and diffusion tensor imaging accounted for variation in clinical decline beyond volume alone. Sample size estimates for atrophy and diffusion imaging were comparable to clinical measures. Corpus callosal fractional anisotropy led to the lowest sample size estimates for all three syndromes. These findings provide further guidance on selection of trial endpoints for studies in frontotemporal dementia and support the use of neuroimaging, particularly structural and diffusion weighted imaging, as biomarkers. Diffusion and perfusion imaging appear to offer additional utility for explaining clinical change beyond the variance explained by volume alone, arguing for considering multimodal imaging in treatment trials.

1 Department of Neurology, Memory and Aging Center, University of California at San Francisco (UCSF), San Francisco, CA, USA
2 Department of Epidemiology and Biostatistics, University of California at San Francisco (UCSF), San Francisco, CA, USA
3 Department of Neurology, College of Medicine, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905, USA
4 Frontotemporal Disorders Unit, Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Charleston, MA, USA

Received August 4, 2017. Revised September 10, 2018. Accepted November 2, 2018

5 Department of Pathology, Memory and Aging Center, University of California at San Francisco (UCSF), San Francisco, CA, USA
6 Department of Pathology - LIM 22, University of Sao Paulo Medical School, Sao Paulo, Brazil

Correspondence to: Adam Staffaroni, PhD
Assistant Professor
University of California, San Francisco (UCSF)
Memory and Aging Center
Department of Neurology
675 Nelson Rising Lane
San Francisco, CA 94143, USA
E-mail: adam.staffaroni@ucsf.edu

**Keywords:** brain atrophy and diffusion imaging; cognition; memory and executive function; biomarkers; primary progressive aphasia

**Abbreviations:** ASL = arterial spin labelling; (bv)FTD = (behavioural variant) frontotemporal dementia; CDR = Clinical Dementia Rating Scale; DTI = diffusion tensor imaging; FAQ = Functional Activities Questionnaire; FTLD = frontotemporal lobar degeneration; LME = linear mixed effects; MMSE = Mini-Mental State Examination; nfv/svPPA = non-fluent/semantic variant of primary progressive aphasia

# Introduction

Neurodegenerative diseases such as Alzheimer's disease and frontotemporal dementia (FTD) are devastating disorders caused by the accumulation and spread of toxic proteins in the brain (Seeley, 2017). Disease-modifying treatments targeted at the underlying proteinopathies are being aggressively developed (Boxer *et al.*, 2013; Panza *et al.*, 2016), and trials of these drugs require the identification of reliable endpoints that track disease-related changes over time. Most pivotal (phase 3) trials to date have focused on Alzheimer's disease, which is caused by the accumulation of intracerebral amyloid-β and tau (Ittner and Götz, 2011). Preclinical studies are beginning to produce promising treatments directed at the biological mechanisms associated with FTD. Clinical trials with these agents face a significant challenge due to the clinical heterogeneity in FTD, which can present as a number of syndromes, including the behavioural variant of frontotemporal dementia (bvFTD), which presents with changes on socio-emotional function, the semantic variant of primary progressive aphasia (svPPA), which is associated with loss of knowledge about words and objects, and the non-fluent variant of PPA (nfvPPA), which presents with articulation difficulties and agrammatism. Pathologically, these disorders are all linked because they are usually caused by accumulation of either tau or TDP-43 (TAR DNA-binding protein 43, encoded by *TARDBP*). For instance, svPPA is usually caused by TDP-43 accumulation, and this is also true for about half of bvFTD patients (Perry and Miller, 2013). The other half of patients with bvFTD have tau pathology, which is also the cause in many cases of nfvPPA. Thus, drugs targeting tau or TDP-43 could benefit patients with a variety of symptoms. Trials directed at tauopathies could potentially enrol some bvFTD and nfvPPA patients, while studies of TDP-43 might include svPPA as well as some bvFTD and nfvPPA patients. Yet, the phenotypic variability across these syndromes presents a barrier to such studies, because it might be difficult to find endpoints relevant to all participants. One strategy to deal with this problem would be to limit each trial to patients with similar symptoms. FTD is relatively uncommon, however, and narrow inclusion criteria would slow enrolment of already difficult to recruit patients, and would limit drug approval only to those patients with the syndrome that was targeted. Moreover, novel clinical trial designs grounded in genetics or biomarkers, such as basket trials that include a variety of clinical syndromes, may have advantages for determining treatment effects (Hyman *et al.*, 2017, 2018). These considerations highlight the importance of examining longitudinal trajectories in markers of disease in clinically heterogeneous populations in order to identify endpoints that would allow reasonable sample sizes in clinically mixed groups. The frontotemporal lobar degeneration neuroimaging initiative (FTLDNI) was begun in 2009 to examine longitudinal changes in bvFTD, svPPA and nfvPPA. In addition to their relevance for drug development in FTD, the lessons learned from studying this clinically heterogeneous group of disorders have value for planning clinical trials in other disorders characterized by clinical variability, including Alzheimer's disease.

There are several considerations in the selection of endpoints for a clinical trial aimed at measuring clinical effects of an intervention. The key characteristic for estimating the sample size that will be required is the effect size for the measure of interest, which is a function of the magnitude and variance of the change over time (Sakpal, 2010). Effect size, however, is not the only consideration. Clinical meaningfulness is also critical in pivotal trials. For example, a circumscribed neuropsychological measure may be quite sensitive to change over time, but improvement on that measure may not have a substantial impact on quality of life, day-to-day function, or caregiver burden. Such a measure could be very helpful in phase 1 or phase 2 trials to provide an indication that a drug could have a clinical impact, but regulators such as the US FDA generally

require pivotal clinical trials to include clinically meaningful endpoints (Katz, 2004; Fleming and Powers, 2012). Typically, measures tracking clinicians' impressions of disease severity (Boxer *et al.*, 2013) or change or activities of daily living (Vandenberghe *et al.*, 2016) have been used for this purpose. Such measures also have the advantage of being equally relevant to patients suffering from a variety of symptoms, and thus have particular value in studies including more than one clinical variant. A third consideration is the reliability of the sample size estimate derived from the group that was studied, which may not always represent the full spectrum of patients, especially if the group is relatively small. Because FTD is uncommon, it is challenging to recruit multiple large groups, and extrapolating to a sample size estimate in an independent group with a different patient composition is therefore difficult. Additional analyses to estimate the likelihood that a given measure will yield a similar effect size in another group can be valuable. In the absence of replicated effect sizes, measures that appear to be more reliable based on the available data would be a sensible choice, and quantification of that reliability can inform the decision about how many patients should be recruited.

Lastly, while clinical improvement is the gold standard for deciding on whether a drug is beneficial for a group of patients, there are reasons to develop other measures. Because many drugs are directed at the underlying protein disorders, they can be considered as potentially disease-modifying therapies (Tsai and Boxer, 2016). Disease-modifying therapies are likely to be most effective at early stages of disease when clinical measures may be less sensitive to change. Brain imaging is an attractive outcome because it provides direct biological information that can be used to support claims of disease modification and can be deployed even in asymptomatic trial participants. Although imaging could be accepted as a surrogate outcome in early phase trials even without a known association with clinical symptoms, establishing reliable relationships between radiographic and clinical measures may increase confidence in selecting imaging as a surrogate endpoint, even in late-phase efficacy trials. Prior studies have demonstrated that brain imaging can be tracked reliably over time in FTD, and the effect sizes can be larger than those observed for clinical measures (Knopman *et al.*, 2009; Pankov *et al.*, 2016). Brain volume has been studied most extensively, and longitudinal change in whole brain volume has been shown to correlate with clinical changes in FTD (Knopman *et al.*, 2009; Gordon *et al.*, 2010). There are several additional candidate imaging biomarkers for tracking change in FTD, including diffusion tensor imaging (DTI) to track white matter integrity, and arterial spin labelling (ASL) to quantify cerebral perfusion. Although previous work suggests these types of imaging are valuable for diagnosis (Du *et al.*, 2006; Santillo *et al.*, 2013; Zhang *et al.*, 2013; Tosun *et al.*, 2016) and white matter integrity can be tracked over time (Lam *et al.*, 2014; Mahoney *et al.*, 2015; Tu *et al.*, 2015), longitudinal studies with these modalities have been limited and correlations between longitudinal changes in DTI and perfusion and clinical measures have not been studied in FTD, nor has the question of whether these images can explain clinical changes beyond what can be accounted for using volume loss.

In this study, we examined longitudinal changes in clinical and imaging markers of disease in bvFTD, svPPA and nfvPPA to identify the potential endpoints for clinical trials for each of these syndromes, as well as those that might be suitable for tracking disease in a study that would include patients in more than one of these groups. We examined neuropsychological and functional measures as well as three imaging-based measures: cortical volume, white matter integrity, and cerebral perfusion. For the imaging metrics, we support their relevance to clinical outcomes by providing evidence of their relationship with clinical measures. Lastly, we provide preliminary sample size estimates for using these clinical and imaging measures in potential clinical trials, and use bootstrapping methods to generate confidence intervals for the sample size estimates.

# Materials and methods

## Participants

Longitudinal data were studied in 161 patients with FTD syndromes (77 bvFTD; 45 svPPA; 39 nfvPPA) and 137 controls. Most were studied through the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI, AG032306), which was funded to examine longitudinal clinical and imaging changes in FTD; some were enrolled in another UCSF FTD study (AG019724) with similar imaging and clinical protocols. We elected to combine data from the two studies to generate the largest dataset possible, which resulted in variability in sample sizes and time points across measures. Further details regarding the number of follow-up visits are provided in Supplementary Table 1 and Supplementary Fig. 1.

Patients were studied at one of three medical centres [University of California, San Francisco (UCSF, $n = 132$), Mayo Clinic, Rochester (MCR, $n = 17$) and Massachusetts General Hospital (MGH, $n = 12$)]. Patients were referred by outside physicians or self-referred, and all underwent neurological, neuropsychological and functional assessment with informant interview. All were diagnosed at a multidisciplinary consensus conference using published criteria: Neary criteria (Neary *et al.*, 1998) or the recently published consensus criteria for bvFTD (Rascovsky *et al.*, 2011) and PPA (Gorno-Tempini *et al.*, 2011) depending on year of enrolment. Demographics and selected clinical features are described in Table 1. The protocol was approved by IRBs at all sites.

The neuropsychological battery has been described elsewhere (Knopman *et al.*, 2008) and included the Mini-Mental State Examination (MMSE; Folstein *et al.*, 1975); a copy of the Benson complex figure (Kramer *et al.*, 2003); to assess visuospatial function, forward and backward digit span (Wechsler, 1997); a modified Trail-making task (Kramer *et al.*, 2003); the Stroop inhibition task (Stroop, 1935); a design fluency task (Delis *et al.*, 2001); a 15-item Boston Naming Test (BNT; Kaplan *et al.*, 1983); phonemic fluency (words beginning

**Table 1 Demographics**

| | BvFTD | SvPPA | NfvPPA | Controls |
|---|---|---|---|---|
| *n* | 77 | 45 | 39 | 137 |
| Average number of visits (range) | 2.7 (1–6) | 2.9 (1–6) | 2.7 (1–6) | 3.2 (1–7) |
| *n* with structural imaging | 55 | 39 | 34 | 137 |
| Average number of imaging visits (range) | 2.8 (1–6) | 3.3 (1–6) | 3.1 (1–6) | 3.1 (1–7) |
| Base age | 62.06 (6.25) | 63.47 (6.19) | 67.67 (7.31) | 63.41 (7.3) |
| Education | 15.62 (3.24) | 16.58 (3.0) | 16.44 (3.3) | 17.47 (1.87) |
| Baseline MMSE | 24.04 (4.51) | 24.88 (4.75) | 25.17 (4.49) | 29.37 (0.79) |
| Baseline FTLD-CDR sum of boxes | 8.25 (4.52) | 5.74 (2.87) | 4.07 (2.5) | 0.04 (0.18) |
| Disease duration at baseline | 5.23 (3.68) | 5.08 (2.44) | 4.29 (3.63) | NA |

NA = not available.

with the letter 'D'/minute) (Kramer *et al.*, 2003); semantic fluency (animals/minute) (Delis *et al.*, 2001); an abbreviated version of the Peabody Picture Vocabulary Test-Revised (PPVT) (Dunn and Dunn, 1981); Pyramids and Palm Trees, a measure of semantic access (Howard and Patterson, 1992); the California Verbal Learning Test, Second Edition-Short Form (CVLT) (Delis *et al.*, 2000); and a test of memory for the Benson figure.

Several measures were used to quantify functional state including the FTLD-modified Clinical Dementia Rating Scale (FTLD-CDR) sum-of-the-boxes score (Morris, 1993; Daly *et al.*, 2000; Knopman *et al.*, 2008), the Functional Activities Questionnaire (FAQ; Pfeffer *et al.*, 1982) the Clinician's Global Impression Scale (CGI; Busner and Targum, 2007) and the Schwab and England Activities of Daily Living scale (SEADL; Schwab and England, 1969). The Neuropsychiatric Inventory (NPI; Cummings *et al.*, 1994) was used to quantify behavioural abnormalities, and the Geriatric Depression Scale was used to measure depressive symptomatology (GDS; Yesavage *et al.*, 1982).

## Neuroimaging acquisition and analysis

Magnetic resonance images at UCSF and MGH were acquired on a 3 T Siemens Tim Trio system equipped with a 12-channel head coil. At MCR, images were acquired on 3 T GE MRI scanner equipped with an 8-channel head coil. Volumetric images from all sites were analysed as one group. Because of the fact that methods for analysis of multisite diffusion and ASL imaging data are less well established, only the UCSF diffusion weighted and ASL images were analysed. Images were also graded for white matter hyperintensity burden using the Fazekas scoring system to examine the potential impact of this measure on the results. Details of image acquisition, processing and grading are included as Supplementary material.

### Biomarkers and pathological analysis

To assess whether the results could be influenced by a large number of patients that might have Alzheimer's disease pathology, we categorized each patient with respect to likely presence or absence of amyloid pathology for whom pathological or Alzheimer's disease biomarker data were available ($n = 109$). The classification was based on autopsy diagnosis if available ($n = 34$), or amyloid imaging results ($n = 53$), or CSF results in those without autopsy or amyloid imaging results ($n = 22$). Further details about these procedures and criteria for classification are presented in Supplementary Fig. 2.

### Neuroimaging composites

To reduce the number of regions examined, we created lobar composites by summing regions within the frontal and temporal lobes for MRI and ASL. To allow comparison of regional changes in DTI with regional changes in volume and ASL, we created composite regions representing tracts emanating from the right and left frontal and temporal lobes, along with the uncinate fasciculus and genu of the corpus callosum. A detailed explanation of the regions of interest included in each composite can be found in the Supplementary material.

## Hypotheses and rationale for data presentation

Based on previous studies, a number of findings were expected. These included that:

(i) Functional measures would exhibit longitudinal declines across all three syndromes and result in relatively large effect sizes for all groups. In contrast, neuropsychological measures with the largest effects sizes are expected to be specific to each disease.
(ii) Neuroimaging measures will result in estimated sample sizes that are consistent with or lower than cognitive and functional measures.
(iii) Baseline volume of the frontal and temporal lobes will be statistically significant predictors of disease trajectory (i.e. rate of decline).
(iv) We expect that structural neuroimaging will show longitudinal associations with clinical and cognitive outcomes. We also anticipate that diffusion imaging of the white matter microstructure and ASL perfusion imaging will explain additional variance in clinical decline above and beyond structural imaging.

The complexity of the dataset, with many outcomes to be examined in three different groups over multiple observations, would allow many options for data presentation and analysis. The choices for this work were driven by the overall goal of providing data to guide planning of clinical trials. Thus, although differences between groups were examined for

statistical significance, the presentation and analysis focused on a thorough description of change in each of the groups, beginning with presentation of measures that show statistically significant change in each group, and identifying variables where baseline values predicted rates of change. Linear mixed effects (LME) models are ideal for these analyses. To illustrate the impact of using these baseline values to guide trial enrolment, we present the change in sample size estimate that results from using a recruitment strategy that would exclude patients least likely to show change for a trial that would target a treatment effect of 40% reduction in rate of decline. These analyses are followed by comprehensive sample size estimates for all variables that show significant change in each group targeting both a small (25%) and moderate (40%) effect, with associated bootstrapped confidence intervals. To emulate data that would be available in a clinical trial, these sample size calculations use estimates of change over approximately 1 year using only two time points. Lastly, to examine relationships between imaging variables and clinical variables, we used LME models, as described below. In all clinical analyses, data on all functional variables are presented because of their inherent clinical meaningfulness, along with MMSE, which is a commonly used measure of cognition, and two measures of cognition chosen because of their relatively low estimates across all cohorts: a semantic fluency task (animals/min) and a measure of single-word comprehension (PPVT-R). The Supplementary material provides additional information on all variables examined.

## Statistical analysis

Full details of the statistical analysis, including model equations, are presented in the Supplementary material. We estimated the annualized longitudinal change in clinical and imaging outcomes by fitting LME models with random intercepts and slopes, using time (in years) from baseline as the predictor. Additional baseline predictors considered included age, education and sex. Total intracranial volume was entered as a covariate in all models that included structural imaging. We examined our data for possible non-linear relationships between outcome and time by adding quadratic time terms to the LME models for a few primary measures (i.e. FTLD-CDR, FAQ and MMSE). In the resulting plots (not shown), it was clear that divergence from linearity occurred at later time points (>2 years) and this appeared to be driven by a few cases with a particularly long duration of follow-up. We therefore focus on the linear representations of results, i.e. in order to avoid risk of over-interpretation. See the Supplementary material for greater detail. Missing data were treated as missing at random.

Separate models were created for each clinical and imaging outcome, with an unstructured covariance matrix to assess whether baseline performance or imaging metrics predict the rate of change, by modelling the correlation between slope and intercept. The associations among neuroimaging (baseline and longitudinal) and clinical outcomes were also assessed with LME models. For this set of analyses that involved comparison of the relationships between different types of time-varying imaging and clinical variables, imaging predictors were converted to z-scores to enhance interpretability (i.e. effects are expressed in terms of changes per standard deviation in the imaging variable). We first fit models with structural imaging as predictors. Subsequently, DTI and ASL were then added into each model. The improvement in fit between models was tested using the likelihood ratio (LR) test, and we estimated the proportion of variance in annualized clinical change explained (in terms of $R^2$) by each brain region (with or without DTI and ASL, depending on the LR test). Similar modelling was conducted to determine whether baseline neuroimaging predicted clinical decline, except that imaging (and its interaction with time) was entered as time-invariant predictor. We report *P*-values without adjusting for multiple comparisons; rather, we use scientific judgement rather than formal methods of adjustment to indicate where caution is warranted despite findings with $P < 0.05$. See Supplementary material for a full description of this decision.

Sample size estimates for hypothetical clinical trials were derived using the annualized change score between baseline and follow-up (Sakpal, 2010). Each estimate was obtained using the *boot* package from 'R' via which a 10 000-fold bootstrapping procedure was performed, allowing the generation of confidence intervals for the estimates. See the Supplementary material for a discussion on error handling. For those measures that showed a statistically significant correlation between baseline values and rate of decline, we assessed the impact of this information on hypothetical clinical trial enrolment. To do this, we removed the 25% of the sample that was least likely to change based on their baseline level for that measure, and recalculated sample size estimates in the remaining patients.

## Data availability

The data that support the findings of this study are openly available in the Laboratory of Neuroimaging (LONI) Image Data Archive at https://ida.loni.usc.edu.

# Results

## Autopsy findings and Alzheimer's disease biomarkers

Among the 34 autopsied patients, all had a primary pathological diagnosis of FTLD. In those with bvFTD ($n = 17$), 10 had FTLD-tau and seven had FTLD-TDP. All three svPPA patients had FTLD-TDP, and among those with nfvPPA ($n = 13$), 12 had FTLD-tau and one had FTLD-TDP. In those with molecular PET imaging ($n = 53$), all were PIB negative. Finally, 3 of 22 cases analysed for CSF data exhibited a total tau/amyloid-β ratio suspicious of Alzheimer's disease pathology (cut-off = 0.39). One of these had a bvFTD motor neuron disease phenotype, strongly suggestive of underlying FTLD-TDP despite the CSF findings.

## Clinical changes and sample size estimates

All the functional measures (FTLD-CDR-SB, FAQ, SEADL and CGI) showed statistically significant longitudinal change for all three diagnoses, as did MMSE. The FTLD-CDR-SB was the only clinical measure to show different

rates of declines between the three diagnoses, with the bvFTD group showing greater change over time than the svPPA group; both groups showed greater change than nfvPPA (Tables 2 and 3). Increases in behavioural disturbance as measured by NPI total severity score only reached statistical significance in the nfvPPA group, but the estimated changes were in the same direction for bvFTD and svPPA. Among neuropsychological variables, semantic fluency and PPVT were among those with the highest effect size for change over time in the LME models in all three groups. Baseline and annual change in scores for all these variables are provided in Tables 2 and 3, as are group differences in rates of decline. Baseline and longitudinal data for all clinical variables, for the three FTD cohorts and controls, are presented in Supplementary Table 3. Figure 1A and B display raw data and the estimated mean change for two common clinical endpoints, FTLD-CDR-SB and MMSE.

For each clinical variable, we have provided sample size estimates and bootstrapped confidence intervals for a theoretical randomized 1:1 placebo controlled study targeting a 25% or 40% reduction in rate of decline from an intervention (power of 0.8), accounting for an anticipated attrition rate of 20% (Supplementary Table 2) and assuming the same standard deviation of changes in the treatment group as that observed for controls. For each diagnosis, the most promising of the 40 measures based on bootstrapped confidence intervals are displayed in Fig. 2. Several functional rating scales and neuropsychological measures have estimates of ~200 or less for both arms, accounting for attrition, to detect a 40% effect at a power of 0.8. In bvFTD these include FTLD-CDR-SB ($n = 178$) and FAQ ($n = 202$); in svPPA, FAQ ($n = 156$), MMSE ($n = 173$) and semantic fluency ($n = 120$), and in nfvPPA, the Stroop color naming task was the best performing neuropsychological endpoint ($n = 26$). The FAQ had the lowest estimated sample size for a functional measure ($n = 240$) in nfvPPA. It is worth noting that there was a substantial overlap in the confidence intervals for all of these measures, with the exception of Stroop color naming in nfvPPA, which was significantly lower than FAQ. Additional sample size calculations for all variables examined are provided in Supplementary Table 2.

There were several variables in each group where a lower value at baseline (or higher in the case of FTLD-CDR-SB) statistically significantly predicted more rapid rate of change over time. In bvFTD, these included FAQ, SEADL and MMSE; in svPPA, the FTLD-CDR-SB, SEADL, and digit span backward; in nfvPPA, FTLD-CDR-SB, MMSE, and Stroop interference. For variables/groups where this effect was statistically significant, we have provided illustrative examples of the effect on sample size estimation if those with baseline scores in the quartile least likely to show change were excluded (Tables 2 and 3, rows labelled 'Enriched Trial'). For example, in bvFTD this would decrease the total sample size from 202 to 113 for a proposed study using the FAQ as an endpoint. Given that 41 estimates were conducted in each group, some findings may have occurred by chance.

## Imaging changes and sample size estimates

All four composite regions of interest showed statistically significant declines in volume for all groups (Table 3). Rapidity of decline in regional volumes differed between groups in the hypothesized pattern. For example, right frontal volume showed faster declines in bvFTD compared to nfvPPA. Fractional anisotropy values also declined over time for all seven composite regions of interest across nearly all diagnoses, again with some *a priori* expected between-group differences. The only DTI region of interest where the observed change did not meet statistical significance (at the nominal 5% level) was the right uncinate fasciculus in nfvPPA, but the change was in the expected direction. Table 3 provides baseline values and change estimates (in cubic centimetres for volume and fractional anisotropy units for DTI) for each of the regions of interest examined, and sample size estimates are presented in Table 4. Sample size estimates at varying levels of effect size are presented in Supplementary Table 2.

Fractional anisotropy in the genu of the corpus callosum consistently produced low sample sizes across syndromes ($n = 94$–$147$ for the trial), although a relatively larger confidence interval was observed in the nfvPPA group ($n = 65$–$797$) compared to the other two diagnoses. The second-best sample size estimates from DTI in each cohort were identified in tracts expected to be involved early and consistently in each syndrome: right uncinate for bvFTD ($n = 168$), left temporal for svPPA ($n = 226$) and left frontal for nfvPPA ($n = 193$). Most ASL regions of interest showed declines as hypothesized, although only four of 12 regions of interest showed statistically significant longitudinal declines; three of these results were observed in the svPPA group. Sample size estimates for ASL were all $>1000$ for a trial, with the exception of left frontal perfusion in nfvPPA ($n = 509$). Figure 1C and D shows raw data and the estimated mean change for two promising volumetric and DTI regions of interest: the left temporal lobe and the genu of the corpus callosum. In a sensitivity analysis, sample size estimates for the entire cohort did not differ substantively from the UCSF cohort (Supplementary Table 4).

We identified one region in svPPA, the right temporal lobe, where baseline volume predicted subsequent rate of change. Specifically, lower volume in svPPA was associated with a relatively slower rate of decline. Consistent with this finding, lower baseline volumes in left and right temporal lobes for all the cohorts were associated with slower rates of decline, although not statistically significant except for the case mentioned above. In contrast, lower baseline frontal lobe volumes were associated with faster rates of decline in the same regions, although none were statistically significant. For DTI, we found that lower levels of baseline

**Table 2** Baseline values, annualized decline and sample size estimates for clinical measures

| | BvFTD | | SvPPA | | NfvPPA | | Between-FTD group differences | Controls | |
|---|---|---|---|---|---|---|---|---|---|
| **Functional measures** | *n* | | *n* | | *n* | | | *n* | |
| **FTLD-CDR** | | | | | | | | | |
| Mean at baseline (SD) | 76 | 8.39 (3.48) | 45 | 5.78 (2.26) | 38 | 3.97 (2.24) | | 137 | 0.04 (0.14) |
| Mean annual decline (SD) | 3.24 (2.75)* | | 1.91 (1.18)* | | 1.55 (1.88)* | | BV > SV > NFV | | −0.005 (0.00) |
| Sample size[a] | 64 | 178 | 40 | 643 | 30 | 463 | | | |
| Enriched sample size | | – | 28 | 605 | 20 | 247 | | | |
| **FAQ** | | | | | | | | | |
| Mean at baseline (SD) | 56 | 18.52 (6.51) | 40 | 9.89 (4.9) | 34 | 6.53 (5.81) | | 137 | 0.30 (0.86) |
| Mean annual decline (SD) | 3.73 (4.45)* | | 4.07 (2.45)* | | 4.45 (3.52)* | | None | | 0.23 (0.33) |
| Sample size[a] | 39 | 201 | 26 | 154 | 20 | 238 | | | |
| Enriched sample size | 30 | 113 | | – | | – | | | |
| **SEADL** | | | | | | | | | |
| Mean at baseline (SD) | 42 | 57.8 (11.85) | 33 | 81.33 (11.15) | 30 | 81.08 (11.7) | | 48 | 100 (0) |
| Mean annual decline (SD) | −10.77 (5.95)* | | −8.67 (7.33)* | | −11.18 (8.0)* | | None | | |
| Sample size[a] | 19 | 410 | 13 | 358 | 13 | 378 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **CGI severity** | | | | | | | | | |
| Mean at baseline (SD) | 62 | 3.84 (0.52) | 39 | 3.29 (0.62) | 33 | 3.58 (0.33) | | 67 | 1.03 (0.07) |
| Mean annual decline (SD) | 0.45 (0.24)* | | 0.6 (0.47)* | | 0.31 (0.24)* | | None | | 0.02 (0.00) |
| Sample size[a] | 30 | 735 | 17 | 197 | 11 | N/A | | | |
| Enriched sample size | | – | | – | | | | | |
| **NPI** | | | | | | | | | |
| Mean at baseline (SD) | 72 | 13.4 (4.16) | 41 | 9.21 (2.6) | 35 | 4.71 (3.14) | | 137 | 0.55 (0.97) |
| Mean annual decline (SD) | 0.14 (1.37) | | 0.11 (0.56) | | 0.84 (0.81)*** | | None | | −0.06 (0.00) |
| Sample size[a] | >2000 | | >2000 | | >2000 | | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Neuropsychological** | | | | | | | | | |
| **MMSE** | | | | | | | | | |
| Mean at baseline (SD) | 76 | 23.82 (4.13) | 44 | 25.14 (4.85) | 37 | 24.44 (5.08) | | 137 | 29.39 (0.40) |
| Mean annual decline (SD) | −3.98 (4.77)* | | −3.75 (2.15)* | | −0.99** | | None | | 0.006 (0.00) |
| Sample size[a] | 62 | 328 | 35 | 172 | 26 | 454 | | | |
| Enriched sample size | 45 | 274 | | – | | 394 | | | |
| **Semantic fluency** | | | | | | | | | |
| Mean at baseline (SD) | 74 | 9.58 (5.31) | 42 | 8.41 (4.34) | 37 | 10.48 (6.07) | | 137 | 23.9 (4.23) |
| Mean annual decline (SD) | −1.77 (1.82)* | | −1.84 (1.04)* | | −1.62 (1.75)* | | None | | −0.15 (0.21) |
| Sample size[a] | 52 | 1501 | 30 | 118 | 23 | 507 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Lexical fluency** | | | | | | | | | |
| Mean at baseline (SD) | 74 | 6.49 (3.56) | 42 | 8.33 (4.0) | 37 | 5.76 (2.98) | | 135 | 16.23 (3.28) |
| Mean annual decline (SD) | | −1.0 (0.64)* | | −1.29 (1.24)* | | −0.83 (0.54)* | None | | 0.33 (0.58) |
| Sample size[a] | 50 | 1317 | 40 | 1041 | 24 | >2000 | | | |
| Enriched sample size | | – | | – | | – | | | |

[a]Sample size estimates calculated for entire trial at power of 0.80 and a 40% effect size, accounting for 20% attrition.

Sample size estimates were calculated in a subsample as described in the 'Materials and methods' section. N/A = bootstrapping did not converge. Enriched sample sizes were reported when the covariance between slope and intercept was significant ($P < 0.05$); otherwise '–' was listed. Sample sizes are not reported for arterial spin labelling (ASL) perfusion due to the large estimates across regions of interest (>500 per arm).

BV = behavioural variant frontotemporal dementia; CGI = Clinician's Global Impression Scale; NFV = non-fluent variant primary progressive aphasia; NPI = Neuropsychiatric Inventory; SEADL = Schwab and England Activities of Daily Living scale; SV = semantic variant PPA.

*$P < 0.001$; **$P < 0.05$; ***$P = 0.039$.

fractional anisotropy in the genu of the corpus callosum predicted a faster rate of decline in fractional anisotropy for the bvFTD and nfvPPA groups. For these regions, we provide in Table 3 sample size estimates for an 'enriched' clinical trial using DTI as an outcome. For example, enrichment based on baseline genu fractional anisotropy in the nfvPPA reduced the sample size for the trial from 147 to 67.

## Neuroimaging correlates of clinical change

We identified statistically significant correlations between imaging metrics and a wide variety of functional and neuropsychological measures. In many cases, likelihood ratio tests demonstrated that the addition of DTI and ASL as predictors significantly improved model fit. Table 5 presents the longitudinal imaging correlates of changes in three commonly used endpoints (i.e. FTLD-CDR, FAQ, MMSE). For these variables, frontal and temporal ASL and fractional anisotropy did not offer statistically significant additional explanatory power above and beyond volume in bvFTD. In svPPA, fractional anisotropy of the left frontal and temporal lobes explained additional variance in change of clinical outcomes, and in nfvPPA, fractional anisotropy and ASL values of several regions explained changes in clinical outcome beyond volume alone. Across the diagnoses, $R^2$ values quantifying the variance in change of FTLD-CDR and FAQ scores explained by imaging ranged from 0.52 to 0.79. Data for all other clinical variables are provided in Supplementary Tables 5–7.

Table 6 presents baseline imaging predictors of the rate of decline in the same three outcomes described in Table 5. A similar pattern was observed for baseline predictors of clinical change, with predictions in the nfvPPA cohort benefitting most frequently from adding ASL and fractional anisotropy to the predictive model. In most cases, worse values on imaging at baseline (e.g. smaller volume) predicted faster rates of clinical decline. Overall, higher model $R^2$ values, in many cases $>0.8$, were observed for cognitive measures compared to clinical measures. Data for all clinical variables are presented in Supplementary Tables 8–10.

# Discussion

The chief aim of this study was to inform clinical trials in FTD by identifying measures that show significant change in each subtype and providing sample size estimates for a range of clinical and neuroimaging measures, and characterize the uncertainty of these estimates using bootstrapping. The results from this analysis suggested that, for trials that might enrol patients with two or three of the canonical FTD syndromes, functional measures appeared be to an attractive clinical outcome measure, because they produce relatively similar sample size estimates for all groups, and have inherent clinical meaningfulness. While previous studies have identified the CDR-SB score as a potentially useful functional measure (Knopman *et al.*, 2008), our analysis suggested that the FAQ is also a good choice. Longitudinal measures of cortical volume showed statistically significant changes over time and generally yielded sample size estimates that were comparable to, and in some cases better than, estimates for clinical measures alone. Changes in white matter integrity also showed statistically significant changes over time and yielded the smallest sample size estimates for brain imaging measures. The upper limits of the confidence intervals were often lower for neuroimaging measures compared to clinical measures, and in many cases, a study powered at the upper limit of the confidence interval would still be feasible. For example, if right temporal lobe volume was used in a svPPA trial powered to detect only a small treatment effect (25% reduction in atrophy), a total sample of 297 would be powered with 95% confidence to detect this effect, even after accounting for an expected attrition of 20%. The confidence intervals were quite revealing for many measures. For example, although the Stroop color naming task was estimated to require a sample size of 287 for a both arms of a bvFTD trial (40% effect size, 0.8 power) accounting for 20% attrition, the upper limit of the 95% confidence interval was 4024, suggesting little confidence that a trial enrolling 287 people would be sufficiently powered to detect this 40% effect size. In contrast, other confidence intervals were much more encouraging.

The second aim was to gauge the suitability of MRI variables as surrogate endpoints in studies attempting to determine either biological effects (i.e. proof of concept or pharmacodynamics) or treatment efficacy. We found that measures of white matter integrity and cortical perfusion, when added to structural imaging, often produced better model fit in explaining variance in clinical changes compared with structural imaging alone. Together, the three types of imaging roughly accounted for between 20 and 95% of the variance in clinical changes. In some cases, changes in volume explained a large percentage of variance in clinical change; for example, in svPPA, volume alone explained 71% of the variance in clinical decline (i.e. FAQ). This strong association provides initial evidence that volumetric MRI may be an appropriate surrogate endpoint for efficacy trials in FTD; this hypothesis can only be confirmed in a successful treatment study. As would be expected, correlations tended to be higher for neuropsychological tasks, which are targeted at specific neural systems, compared with functional measures that depend on combinations of cognitive, behavioural and lifestyle factors. These findings support the large body of accumulated evidence, mostly cross-sectional, correlating imaging measures with clinical findings (Gorno-Tempini *et al.*, 2004; Possin *et al.*, 2012; Perry and Rosen, 2016; Staffaroni *et al.*, 2017). In addition to their potential value in symptomatic FTD, MRI-based measures will become increasingly important in tracking the effects of treatments and predicting

**Table 3 Baseline values, annualized decline and sample size estimates for imaging measures**

| | BvFTD | | SvPPA | | NfvPPA | | Between-FTD group differences | Controls | |
|---|---|---|---|---|---|---|---|---|---|
| **Volumetric data, $cm^3$** | *n* | | *n* | | *n* | | | *n* | |
| **Right frontal** | | | | | | | | | |
| Mean at baseline (SD) | 68 | 35.54 (8.01) | 39 | 42.21 (5.58) | 35 | 39.16 (5.24) | | 137 | 44.68 (4.38) |
| Mean annual decline (SD) | −1.84 (1.33)* | | −1.63 (0.85)* | | −1.18 (0.95)* | | BV > NFV | | −0.26 (0.13) |
| Sample size[a] | 54 | 167 | 34 | 110 | 29 | 365 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Left frontal** | | | | | | | | | |
| Mean at baseline (SD) | 68 | 37.06 (7.2) | 39 | 42.05 (4.92) | 35 | 38.53 (5.75) | | 137 | 45.58 (4.53) |
| Mean annual decline (SD) | −2.08 (1.63)* | | −1.82 (1.03)* | | −1.57 (1.31)* | | None | | −0.25 (0.10) |
| Sample size[a] | 54 | 163 | 34 | 119 | 29 | 303 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Right temporal** | | | | | | | | | |
| Mean at baseline (SD) | 68 | 21.56 (3.77) | 39 | 19.79 (3.11) | 35 | 22.75 (2.82) | | 137 | 24.68 (2.49) |
| Mean annual decline (SD) | −0.85 (0.81)* | | −1.11 (0.46)* | | −0.43 (0.44)* | | (SV and BV) > NFV | | −0.15 (0.06) |
| Sample size[a] | 54 | 240 | 34 | 62 | 29 | 549 | | | |
| Enriched sample size | | – | 25 | 74 | | – | | | |
| **Left temporal** | | | | | | | | | |
| Mean at baseline (SD) | 68 | 22.64 (3.58) | 39 | 18.18 (2.45) | 35 | 22.85 (2.93) | | 137 | 25.47 (2.62) |
| Mean annual decline (SD) | −0.91 (0.87)* | | −1.0 (0.36)* | | −0.64 (0.55)* | | SV > NFV | | −0.14 (0.05) |
| Sample size[a] | 54 | 224 | 34 | 71 | 29 | 399 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **DTI: fractional anisotropy** | | | | | | | | | |
| **Right frontal** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 0.38 (0.04) | 34 | 0.41 (0.03) | 35 | 0.39 (0.03) | | 103 | 41.5 (0.02) |
| Mean annual decline (SD) | $-8.2 \times 10^{-3}$ (0.006)* | | $-3.6 \times 10^{-3}$ (0.002)** | | $-4.6 \times 10^{-3}$ (0.002)* | | None | | $-1.1 \times 10^{-3}$ (0.00) |
| Sample size[a] | | 503 | | >1000 | | 784 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Left frontal** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 0.38 (0.04) | 34 | 0.40 (0.01) | 35 | 0.39 (0.03) | | 103 | 0.42 (0.02) |
| Mean annual decline (SD) | $-8.5 \times 10^{-3}$ (0.002)* | | $-9.1 \times 10^{-3}$ (0.002)* | | −0.01 (0.005)* | | None | | $-9.5 \times 10^{-5}$ (0.00) |
| Sample size[a] | | 458 | | 762 | | 193 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Right temporal** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 37.5 (0.03) | 34 | 0.39 (0.04) | 35 | 0.39 (0.03) | | 103 | 0.40 (0.03) |
| Mean annual decline (SD) | $-4.7 \times 10^{-3}$ (0.004)** | | $-9.7 \times 10^{-3}$ (0.002)* | | $-6.7 \times 10^{-3}$ (0.004)* | | SV > BV | | $-2.1 \times 10^{-3}$ (0.00) |
| Sample size[a] | | 1079 | | 889 | | 1297 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Left temporal** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 0.38 (0.03) | 34 | 0.37 (0.02) | 35 | 0.39 (0.02) | | 103 | 0.40 (0.03) |
| Mean annual decline (SD) | $-6.7 \times 10^{-3}$ (0.005)* | | $-1.4 \times 10^{-2}$ (0.007)* | | $-8.2 \times 10^{-3}$ (0.005)* | | SV > (BV and NFV) | | $-3.8 \times 10^{-3}$ (0.00) |
| Sample size[a] | | 407 | | 226 | | 353 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Right uncinate** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 0.38 (0.01) | 34 | 0.37 (0.07) | 35 | 0.43 (0.05) | | 103 | 0.40 (0.03) |
| Mean annual decline (SD) | $-1.4 \times 10^{-2}$ (0.01)* | | $-1.7 \times 10^{-2}$ (0.01)* | | $-3.1 \times 10^{-3}$ (0.006) | | (BV and SV) > NFV | | $-2.8 \times 10^{-3}$ (0.00) |
| Sample size[a] | | 168 | | 1279 | | 3525 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Left uncinate** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 0.38 (0.06) | 34 | 0.33 (0.04) | 35 | 0.41 (−0.05) | | 103 | 0.39 (0.03) |
| Mean annual decline (SD) | $-1.0 \times 10^{-2}$ (0.01)** | | $-1.5 \times 10^{-2}$ (0.003)* | | $-1.0 \times 10^{-2}$ (0.01)** | | None | | $-2.1 \times 10^{-3}$ (0.00) |
| Sample size[a] | | 309 | | 506 | | 355 | | | |
| Enriched sample size | | – | | – | | – | | | |
| **Genu of corpus callosum** | | | | | | | | | |
| Mean at baseline (SD) | 49 | 0.44 (0.07) | 34 | 0.53 (0.05) | 35 | 0.51 (0.05) | | 103 | 0.57 (0.03) |
| Mean annual decline (SD) | $-2.9 \times 10^{-2}$ (0.02)* | | $-1.7 \times 10^{-2}$ (0.009)* | | $-1.9 \times 10^{-2}$ (0.01)* | | BV > (SV and NFV) | | $-3.7 \times 10^{-3}$ (0.00) |
| Sample size[a] | 24 | 94 | 17 | 101 | 19 | 147 | | | |
| Enriched sample size | 16 | 58 | | – | 14 | 67 | | | |
| **ASL perfusion** | | | | | | | | | |
| **Right frontal** | | | | | | | | | |
| Mean at baseline (SD) | 37 | 16.66 (4.93) | 30 | 22.86 (3.41) | 27 | 20.92 (5.01) | | 68 | 24.1 (5.06) |
| Mean annual decline (SD) | −1.15 (0.67) | | −1.16 (0.1)** | | | 0.27 (0.89) | SV > NFV | | −0.37 (0.38) |
| **Left frontal** | | | | | | | | | |
| Mean at baseline (SD) | 37 | 17.79 (5.32) | 30 | 22.72 (3.17) | 27 | 19.51 (4.39) | | 68 | 24.68 (5.49) |
| Mean annual decline (SD) | −1.13 (0.77) | | −1.2 (0.38)** | | −0.37 (0.73) | | None | | −0.46 (0.65) |

(continued)

**Table 3 Continued**

| | BvFTD | | SvPPA | | NfvPPA | | Between-FTD group differences | Controls | |
|---|---|---|---|---|---|---|---|---|---|
| **Right temporal** | | | | | | | | | |
| Mean at baseline (SD) | 37 | 27.53 (5.14) | 30 | 26.96 (2.26) | 27 | 29.53 (5.21) | | 68 | 29.94 (5.72) |
| Mean annual decline (SD) | −1.44 (0.31)** | | −1.11 (0.24)** | | 0.24 (0.75) | | None | | −0.36 (0.64) |
| **Left temporal** | | | | | | | | | |
| Mean at baseline (SD) | 37 | 27.45 (4.22) | 30 | 24.33 (3.99) | 27 | 27.13 (5.41) | | 68 | 30.41 (5.69) |
| Mean annual decline (SD) | −0.97 (2.37) | | −0.51 (0.24) | | −0.27 (1.40) | | None | | −0.34 (0.00) |

[a]Sample size estimates calculated for entire trial at power of 0.80 and a 40% effect size, accounting for 20% attrition.
Sample size estimates were calculated in a subsample as described in the 'Materials and methods' section. Variant-specific measures represent those three measures (not including MMSE or fluency) with the lowest sample size per variant. Enriched sample sizes were reported when the covariance between slope and intercept was significant ($P < 0.05$); otherwise '–' was listed. Sample sizes are not reported for ASL perfusion due to the large estimates across regions of interest (>500 per arm).
*$P < 0.001$; **$P < 0.05$.

progression in asymptomatic or very mildly symptomatic disease (Rohrer *et al.*, 2015; Lee *et al.*, 2017).

For 17% of all clinical and imaging metrics analysed, we found a statistically significant relationship between baseline performance and rate of change. Baseline imaging was also found to significantly predict the subsequent decline for many clinical variables. Baseline clinical and imaging measures can thus be used to stratify FTD treatment trials. For several measures, those who begin with the lowest level of impairment are the least likely to show change. Stratifying those participants could potentially reduce the sample size necessary to detect an effect. For bvFTD in particular, it has been observed that some patients show very little progression over time. These patients have sometimes been labelled as having a 'phenocopy' syndrome, indicating that they may not have FTLD pathology (Khan *et al.*, 2012), and excluding such patients from clinical trials would help establish efficacy. For illustrative purposes, excluding those bvFTD participants with the least severe functional deficits as measured by the FAQ would result in a reduction of total sample size from 202 to 113. The benefit of this enrichment approach would need to be weighed against the concern that treatments initiated too late in disease might not be effective, but the fact that these methods help to identify patients with larger, more reliable rates of change indicates that there is still an opportunity to intervene in these patients. Similar inclusion criteria have been written into Alzheimer's disease clinical trials (NCT02477800, NCT02880956). Imaging has also been used in studies of Alzheimer's disease to predict individual rates of decline as a strategy for enrichment of clinical trials, including using multiple measures in machine learning algorithms (Zu *et al.*, 2016). Our data suggest that similar approaches might be taken in FTD. An alternative method for factoring in the relationship between baseline performance or volume on the expected rate of decline consideration could be to enrol a larger variety of patients but include modelling baseline performance or volume into the power calculations.

We were able to examine multiple imaging measures for tracking change. As expected, lobar cortical volumes showed reliable changes over time and were consistently among the measures requiring the smallest sample sizes, although not always smaller than those for clinical changes. The benefits relative to clinical changes were highest in svPPA, likely due to the relative clinical and pathological homogeneity of svPPA, which is associated with similar patterns of anatomical involvement across patients (Binney *et al.*, 2017) and with TDP-43 type C pathology in roughly 80% of cases (Spinelli *et al.*, 2016). Changes in white matter integrity also showed reliable change, with sample size estimates that were generally comparable to those seen with structural imaging based on confidence intervals. One measure that stood out was the genu of the corpus callosum, which generated small sample size estimates in all three variants, although it is worth noting that the confidence intervals were wider with a higher upper limit in the nfvPPA cohort. The agreement of this finding across three separate cohorts further validates its utility. Corpus callosum may be a particularly valuable indicator because it carries fibres mediating communication between large parts of the cerebral hemispheres. In addition, DTI, which tracks the movement of water along white matter tracks, is sensitive to the presence of crossing fibres (Lee *et al.*, 2015), and the highly organized nature of fibres passing through the corpus callosum fibres may make this region ideal for detecting subtle changes in tract integrity. Very few studies have examined longitudinal changes in ASL perfusion in FTD, with the exception of a recent study in *GRN* and *MAPT* mutation carriers, which found decreased perfusion over time (Dopper *et al.*, 2016). Although we were able to detect statistically significant decline in volume-corrected perfusion in svPPA, our findings suggest that ASL may not be an ideal method for tracking disease once the signal is corrected for atrophy, at least when using the acquisition and processing methods we studied.

The present study also adds to our knowledge on the trajectory of functional and neuropsychological measures in bvFTD, where clinical presentation and pathology are particularly heterogeneous. Despite the core symptom being behavioural dysfunction, which is captured by the Neuropsychiatric Inventory (NPI), our findings confirm prior studies indicating that the NPI does not provide a

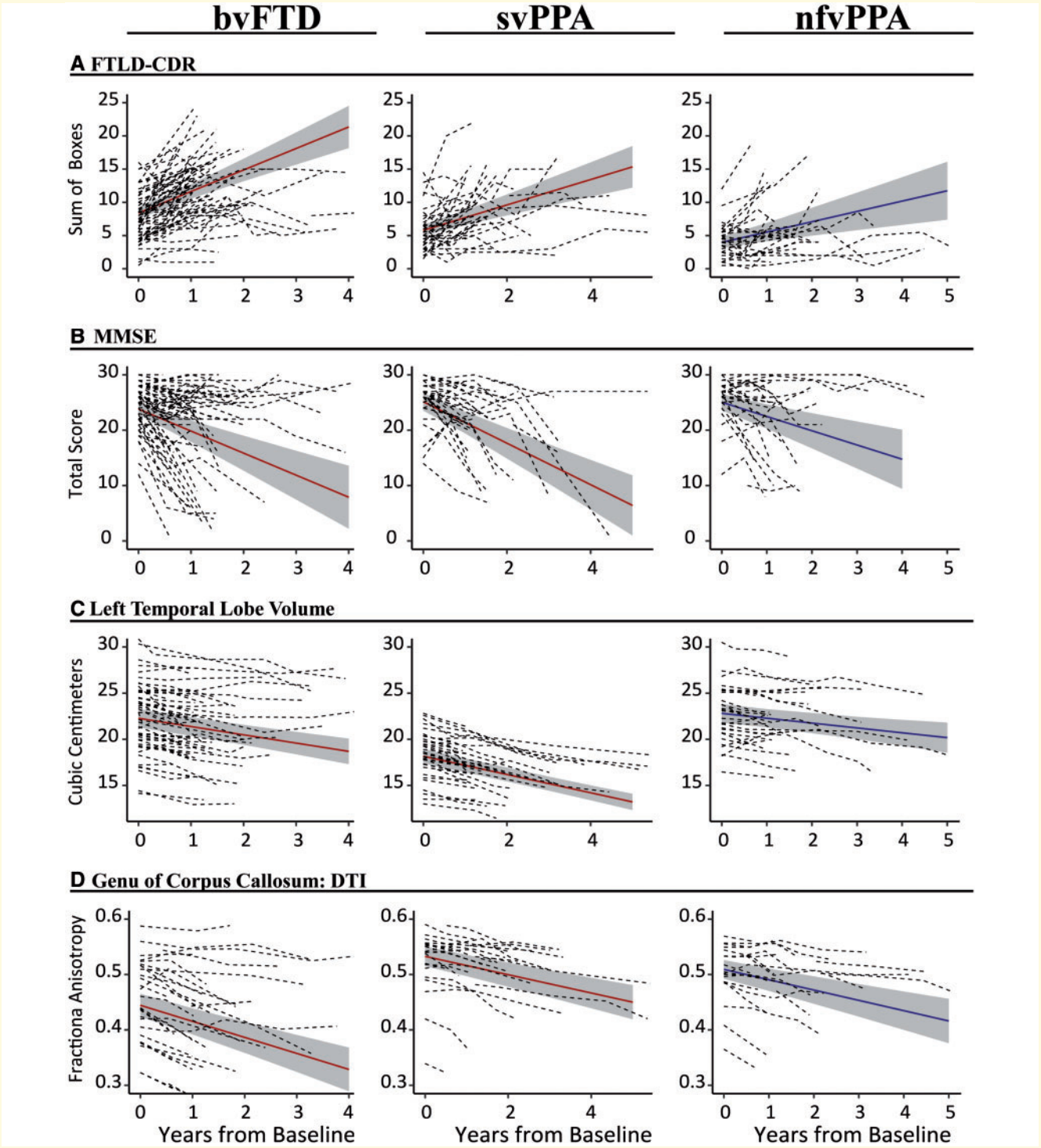


**Figure 1 Longitudinal clinical and imaging outcomes in FTD clinical syndromes.** Raw longitudinal data for all participants are shown in grey lines. Overlaid in colour are the estimated slopes and 95% confidence intervals (in grey) from linear mixed effects models. The FTLD-CDR sum of box score and MMSE total score are shown for all groups given their precedence for use in clinical trials. We also show the trajectory of left temporal lobe volume and corpus callosum fractional anisotropy, which showed promising sample size estimates across all diagnoses.

good measure for clinical trials in this syndrome (Knopman *et al.*, 2008). This likely results from a combination of factors, including the fact that bvFTD patients may already be at the peak of their behavioural difficulties when they present to academic centres, and the fact that the specific behavioural phenotype varies across individuals (Ranasinghe *et al.*, 2016). Our results confirm previous work (Knopman *et al.*, 2008; Schubert *et al.*, 2016;

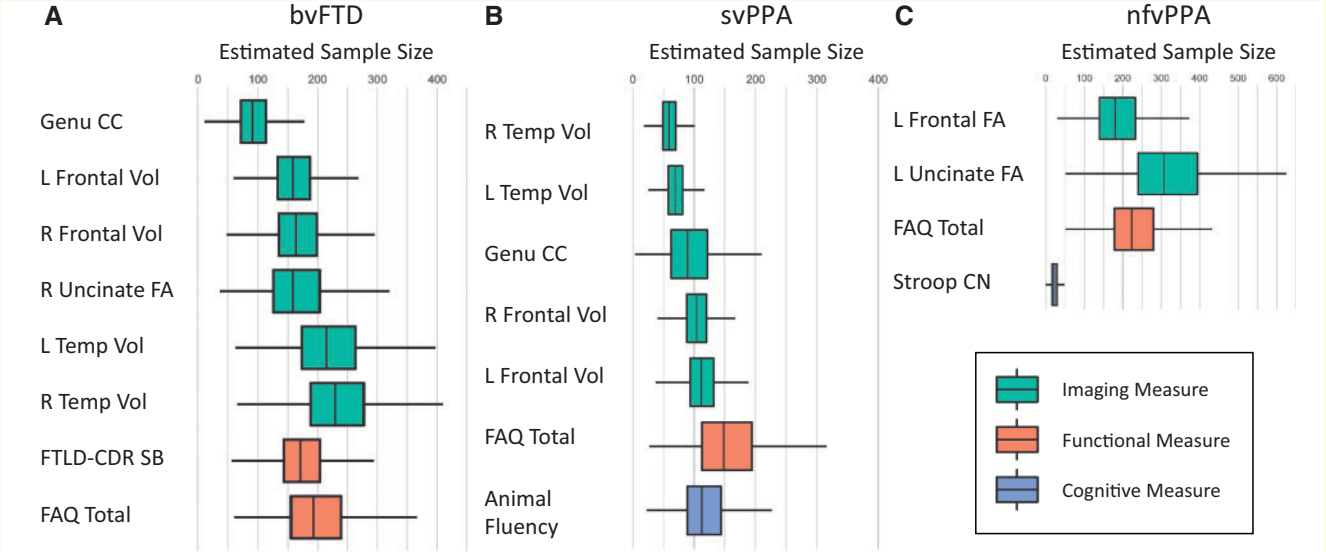


**Figure 2 Estimated sample sizes and bootstrapped confidence intervals.** Sample sizes were estimated for both arms of a randomized control trial powered at 0.8, alpha = 0.05, with an expected attrition of 20%. Confidence intervals (95%) were obtained from a 10 000-fold bootstrap procedure and presented using black bars. For bvFTD and svPPA, measures were included if the upper limit of the confidence interval was $<500$; for nfvPPA, inclusion was extended to measures with an upper limit $<800$. CC = corpus callosum; FA = fractional anisotropy; FAQ = Functional Activities Questionnaire; L = Left; R = Right; Stroop CN = Stroop Color Naming; Temp = Temporal; Unc. = Uncinate Fasciculus; Vol = Volume.

Ramanan *et al.*, 2017); however, showing that bvFTD patients experience reliable decline in many domains of cognition, and that functional measures currently provide the best metric for tracking change in this syndrome.

In contrast to bvFTD, the defining deficits in svPPA and nfvPPA involve loss of speech and language functions. Although tasks tapping semantics (such as naming and word recognition) declined significantly in svPPA and correlated well with left temporal lobe integrity, the clinical measures with the greatest effects sizes in both svPPA and nfvPPA were those having a frontal component (e.g. fluency in svPPA, Stroop, fluency and digit span in nfvPPA) and those tapping general cognitive functions (MMSE and functional scales). These findings suggest that, at a time when syndrome-specific deficits may have advanced considerably and variably in these disorders, the best clinical measures for tracking disease may be those that indicate spread to new regions and have relevance to each individual regardless of their exact clinical deficit. The sample size estimates for svPPA in this study, as well as in previous studies that included svPPA (Knopman *et al.*, 2008; Binney *et al.*, 2017), were among the smallest, again likely due to the clinical and pathological homogeneity. svPPA was the only syndrome where those with the lowest volumes in the group were least likely to show more temporal lobe volume loss, consistent with prior observations indicating that temporal lobe atrophy in svPPA is so profound that it can reach a floor beyond which further volume loss is greatly attenuated (Binney *et al.*, 2017). The Stroop color naming task, which requires speeded naming of blocks of colours, was suggested to be an excellent outcome for tracking clinical change in nfvPPA. The promise of this measure is likely because this task tracks speeded verbal output with the smallest cognitive load, making it a relatively pure indicator of changes in motor speech in this group. It is important to note that the relative differences in sample size estimates reported in this manuscript should not be mistaken for statistical evidence of the superiority of one measure over another (i.e. in terms of statistical significance), and we recommend replication of these results. We did not directly compare effect sizes via statistical tests due to the relatively small sample sizes.

How can investigators use these findings for planning clinical trials? We believe that our findings can inform both the selection of pharmacodynamic biomarkers for early-stage trials, as well as surrogate outcomes for pivotal late-stage studies. In early-stage trials (phase 1 and 2), our data help to identify measures with relatively large and reliable effect sizes, mostly imaging-based measures and cognitive tasks targeted at specific domains, which can provide evidence of biological proof-of-concept in small studies. For such early studies, it might help to increase power by adjusting inclusion criteria using baseline characteristics to include those with a higher likelihood of progression. Our data provide guidance on selecting functional outcome measures with inherent clinical meaning for pivotal phase 3 studies, which require clinically meaningful endpoints. Furthermore, the large effect sizes observed for volumetric imaging, in conjunction with the large degree of variance in clinical function explained by these imaging variables (e.g. $R^2 = 71\%$), provide initial evidence that

**Table 4 Estimated sample sizes and bootstrapped confidence intervals for select measures**

| | | BvFTD | | | SvPPA | | | NfvPPA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Endpoint** | **Effect size, %** | ***n*** | **Sample size** | **95% CI** | ***n*** | **Sample size** | **95% CI** | ***n*** | **Sample size** | **95% CI** |
| **Functional measures** | | | | | | | | | | |
| FTLD-CDR | 40 | 64 | 178 | 115–348 | 40 | 643 | 210–>2000 | 30 | 463 | 211–>2000 |
| | 25 | | 456 | 296–852 | | 1648 | 510–>2000 | | 1186 | 523–>2000 |
| FAQ | 40 | 39 | 201 | 118–451 | 26 | 154 | 77–418 | 20 | 238 | 132–509 |
| | 25 | | 515 | 298–1163 | | 396 | 199–1058 | | 611 | 337–1295 |
| SEADL | 40 | 19 | 410 | 176–>2000 | 13 | 358 | 169–>2000 | 13 | 378 | 204–>2000 |
| | 25 | | 1051 | 455–>2000 | | 917 | 442–>2000 | | 969 | 525–>2000 |
| CGI-Severity | 40 | 30 | 735 | 241–>2000 | 17 | 197 | 66–891 | | – | – |
| | 25 | | 1882 | 599–>2000 | | 505 | 176–>2000 | | – | – |
| **Cognitive measures** | | | | | | | | | | |
| MMSE | 40 | 62 | 328 | 198–785 | 35 | 172 | 86–1103 | 26 | 454 | 231–>2000 |
| | 25 | | 840 | 505–1993 | | 442 | 214–>2000 | | 1163 | 624–>2000 |
| Animal fluency | 40 | 52 | 1501 | 295–>2000 | 30 | 118 | 66–318 | 23 | 507 | 231–>2000 |
| | 25 | | >2000 | 816–>2000 | | 304 | 170–831 | | 1298 | 599–>2000 |
| PPVT | 40 | 33 | 471 | 282–1360 | 23 | 240 | 81–1720 | 23 | 412 | 178–1374 |
| | 25 | | 1207 | 717–>2000 | | 614 | 206–>2000 | | 1055 | 454–>2000 |
| **Neuroimaging measures** | | | | | | | | | | |
| **Structural MRI volume** | | | | | | | | | | |
| Left frontal | 40 | 54 | 163 | 102–292 | 34 | 119 | 74–195 | 29 | 303 | 114–>2000 |
| | 25 | | 418 | 268–747 | | 305 | 192–505 | | 777 | 292–>2000 |
| Right frontal | 40 | 54 | 167 | 102–326 | 34 | 110 | 70–171 | 29 | 365 | 131–1838 |
| | 25 | | 429 | 266–838 | | 283 | 180–436 | | 935 | 347–>2000 |
| Left temporal | 40 | 54 | 224 | 131–457 | 34 | 71 | 45–123 | 29 | 399 | 154–>2000 |
| | 25 | | 573 | 335–1156 | | 183 | 117–316 | | 1021 | 382–>2000 |
| Right temporal | 40 | 54 | 240 | 144–442 | 34 | 62 | 40–115 | 29 | 549 | 127–>2000 |
| | 25 | | 615 | 373–1177 | | 158 | 103–297 | | 1407 | 344–>2000 |
| **DTI (FA)** | | | | | | | | | | |
| Left frontal | 40 | 24 | 458 | 126–>2000 | 17 | 762 | 150–>2000 | 19 | 193 | 94–460 |
| | 25 | | 1172 | 322–>2000 | | 1952 | 424–>2000 | | 495 | 247–1162 |
| Right frontal | 40 | 24 | 503 | 203–>2000 | 17 | >2000 | 912–>2000 | 19 | 784 | 200–>2000 |
| | 25 | | 1287 | 511–>2000 | | >2000 | >2000–>2000 | | >2000 | 494–>2000 |
| Left temporal | 40 | 24 | 407 | 130–>2000 | 17 | 226 | 70–1586 | 19 | 353 | 114–>2000 |
| | 25 | | 1043 | 339–>2000 | | 580 | 191–>2000 | | 904 | 293–>2000 |
| Right temporal | 40 | 24 | 1079 | 158–>2000 | 17 | 889 | 142–>2000 | 19 | 1297 | 192–>2000 |
| | 25 | | >2000 | 413–>2000 | | >2000 | 365–>2000 | | >2000 | 478–>2000 |
| Left uncinate | 40 | 24 | 309 | 136–1472 | 17 | 506 | 110–>2000 | 19 | 355 | 199–1055 |
| | 25 | | 791 | 359–>2000 | | 1295 | 265–>2000 | | 910 | 496–>2000 |
| Right uncinate | 40 | 24 | 168 | 94–498 | 17 | 1279 | 82–>2000 | 19 | >2000 | 210–>2000 |
| | 25 | | 431 | 239–1229 | | >2000 | 200–>2000 | | >2000 | 522–>2000 |
| Corpus callosum | 40 | 24 | 94 | 51–214 | 17 | 101 | 42–334 | 19 | 147 | 65–797 |
| | 25 | | 242 | 131–553 | | 260 | 104–837 | | 376 | 179–>2000 |

Estimated total sample sizes (both arms) for a trial powered at 0.8, alpha = 0.05, with expected attrition of 20%.
Confidence intervals were obtained from a 10 000-fold bootstrap procedure.
Estimates from models that failed to converge are not provided, nor are estimates >2000.
CI = confidence interval; CGI = Clinician's Global Impression Scale; FA = fractional anisotropy; PPVT = Peabody Picture Vocabulary Test-Revised; SEADL = Schwab and England Activities of Daily Living Scale.

neuroimaging may have a place as a surrogate outcome in these late-stage trials. Although replication of our findings in independent datasets would be ideal, this will be difficult to achieve given the prevalence of FTD, and for definitive studies it may be prudent to consider the confidence intervals for the sample size estimates and plan enrolment close to these limits to increase the likelihood that an effect would be detected.

Some limitations of our study should be noted. In order to present a comprehensive view of the changes in clinical measures and multiple types of imaging and still maintain tractability for statistical analysis and readability, we chose only four large regions of interest and limited our analysis of DTI to fractional anisotropy. Future studies should examine potential differences in the utility of various DTI-based measures. In addition, more granular, data-driven analyses

**Table 5 Neuroimaging and clinical measures correlate longitudinally in FTD**

| | Cognitive measures | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FTLD-CDR | | | | FAQ Total | | | | MMSE | | | |
| | Imaging predictor | *b* | *P* | 95% CI | Imaging predictor | *b* | *P* | 95% CI | Imaging predictor | *b* | *P* | 95% CI |
| **BvFTD** | | | | | | | | | | | | |
| Right frontal | Volume | −1.08 | 0.04 | −2.09, −0.06 | Volume | −1.04 | 0.31 | −3.06, 0.97 | Volume | 1.94 | 0.05 | 0, 3.87 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.52 | | | | Model $R^2$ = 0.62 | | | | Model $R^2$ = 0.75 | | | |
| Left frontal | Volume | −1.25 | 0.01 | −2.17, −0.32 | Volume | −1.28 | 0.18 | −3.15, 0.59 | Volume | 1.90 | 0.03 | 0.19, 3.62 |
| | ASL | – | – | – | ASL | – | – | – | ASL | −0.26 | 0.66 | −1.44, 0.91 |
| | FA | – | – | – | FA | – | – | – | FA | 3.51 | 0.00 | 2.08, 4.94 |
| | Model $R^2$ = 0.56 | | | | Model $R^2$ = 0.67 | | | | Model $R^2$ = 0.74 | | | |
| Right temporal | Volume | −1.31 | 0.03 | −2.52, −0.1 | Volume | −1.62 | 0.20 | −4.06, 0.83 | Volume | 1.61 | 0.18 | −0.72, 3.94 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.55 | | | | Model $R^2$ = 0.68 | | | | Model $R^2$ = 0.7 | | | |
| Left temporal | Volume | −1.01 | 0.08 | −2.15, 0.12 | Volume | −1.03 | 0.34 | −3.17, 1.11 | Volume | 3.26 | 0.00 | 1.2, 5.31 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.52 | | | | Model $R^2$ = 0.61 | | | | Model $R^2$ = 0.73 | | | |
| **SvPPA** | | | | | | | | | | | | |
| Right frontal | Volume | −1.86 | 0.02 | −3.4, −0.32 | Volume | −7.14 | 0.00 | −11.85, −2.43 | Volume | 0.70 | 0.60 | −1.92, 3.31 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.53 | | | | Model $R^2$ = 0.71 | | | | Model $R^2$ = 0.29 | | | |
| Left frontal | Volume | −1.27 | 0.15 | −2.99, 0.45 | Volume | −4.43 | 0.04 | −8.61, −0.26 | Volume | 2.10 | 0.15 | −0.79, 4.99 |
| | ASL | 0.21 | 0.63 | −0.64, 1.06 | ASL | – | – | – | ASL | 0.53 | 0.42 | −0.75, 1.81 |
| | FA | −1.25 | 0.02 | −2.27, −0.22 | FA | – | – | – | FA | 4.20 | 0.00 | 2.58, 5.82 |
| | Model $R^2$ = 0.66 | | | | Model $R^2$ = 0.71 | | | | Model $R^2$ = 0.76 | | | |
| Right temporal | Volume | −1.28 | 0.04 | −2.52, −0.04 | Volume | −4.17 | 0.01 | −7.36, −0.98 | Volume | 3.55 | 0.02 | 0.55, 6.55 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.54 | | | | Model $R^2$ = 0.64 | | | | Model $R^2$ = 0.78 | | | |
| Left temporal | Volume | 1.42 | 0.23 | −0.89, 3.72 | Volume | 5.95 | 0.05 | 0.12, 11.77 | Volume | 4.80 | 0.01 | 0.99, 8.6 |
| | ASL | 0.37 | 0.35 | −0.4, 1.13 | ASL | −0.53 | 0.66 | −2.89, 1.83 | ASL | 0.00 | 0.99 | −1.07, 1.08 |
| | FA | −1.13 | 0.01 | −1.97, −0.3 | FA | −3.89 | 0.00 | −6.39, −1.39 | FA | 1.84 | 0.01 | 0.44, 3.24 |
| | Model $R^2$ = 0.67 | | | | Model $R^2$ = 0.54 | | | | Model $R^2$ = 0.83 | | | |
| **NfvPPA** | | | | | | | | | | | | |
| Right frontal | Volume | −0.39 | 0.63 | −1.95, 1.18 | Volume | −2.17 | 0.27 | −6.04, 1.7 | Volume | 3.44 | 0.08 | −0.43, 7.31 |
| | ASL | −0.38 | 0.32 | −1.12, 0.36 | ASL | −0.12 | 0.92 | −2.53, 2.28 | ASL | – | – | – |
| | FA | −1.71 | 0.00 | −2.54, −0.89 | FA | −3.94 | 0.00 | −6.51, −1.37 | FA | – | – | – |
| | Model $R^2$ = 0.74 | | | | Model $R^2$ = 0.65 | | | | Model $R^2$ = 0.88 | | | |
| Left frontal | Volume | 0.27 | 0.70 | −1.12, 1.66 | Volume | −0.74 | 0.65 | −3.92, 2.44 | Volume | 3.44 | 0.03 | 0.27, 6.62 |
| | ASL | −0.43 | 0.23 | −1.11, 0.26 | ASL | −1.15 | 0.25 | −3.1, 0.8 | ASL | – | – | – |
| | FA | −1.72 | 0.00 | −2.58, −0.86 | FA | −3.30 | 0.00 | −5.39, −1.21 | FA | – | – | – |
| | Model $R^2$ = 0.79 | | | | Model $R^2$ = 0.62 | | | | Model $R^2$ = 0.85 | | | |
| Right temporal | Volume | −0.49 | 0.65 | −2.6, 1.61 | Volume | −1.59 | 0.52 | −6.49, 3.3 | Volume | 3.77 | 0.07 | −0.23, 7.78 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.7 | | | | Model $R^2$ = 0.51 | | | | Model $R^2$ = 0.82 | | | |
| Left temporal | Volume | −0.67 | 0.37 | −2.14, 0.81 | Volume | 0.45 | 0.83 | −3.61, 4.51 | Volume | 3.34 | 0.02 | 0.56, 6.13 |
| | ASL | – | – | – | ASL | −0.39 | 0.72 | −2.53, 1.75 | ASL | 0.37 | 0.45 | −0.59, 1.32 |
| | FA | – | – | – | FA | −3.13 | 0.02 | −5.83, −0.42 | FA | 2.02 | 0.03 | 0.19, 3.86 |
| | Model $R^2$ = 0.74 | | | | Model $R^2$ = 0.54 | | | | Model $R^2$ = 0.81 | | | |

A model with only volume as an imaging predictor was compared to a model with all three imaging modalities. If the likelihood ratio test was $<0.05$, suggesting that the model with all three imaging predictors fit better, parameters for this model are shown. Otherwise, the parameters for the volume model are presented. Imaging predictors were standardized; each parameter indicates the units of raw change in the outcome per SD decline in the imaging metric.

CI = confidence interval; FA = fractional anisotropy.

**Table 6 Baseline imaging metrics predict clinical decline in FTD**

| | Cognitive measures | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FTLD–CDR | | | | FAQ Total | | | | MMSE | | | |
| | Imaging predictor | *b* | *P* | 95% CI | Imaging predictor | *b* | *P* | 95% CI | Imaging predictor | *b* | *P* | 95% CI |
| **BvFTD** | | | | | | | | | | | | |
| Right frontal | Volume | −1.38 | 0.00 | −2.04, −0.72 | Volume | −2.77 | 0.00 | −4.33, −1.2 | Volume | 1.48 | 0.19 | −0.74, 3.71 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.78 | | | | Model $R^2$ = 0.74 | | | | Model $R^2$ = 0.97 | | | |
| Left frontal | Volume | −1.02 | 0.00 | −1.72, −0.32 | Volume | −1.66 | 0.03 | −3.16, −0.16 | Volume | −0.44 | 0.76 | −3.29, 2.4 |
| | ASL | – | – | – | ASL | – | – | – | ASL | 2.63 | 0.10 | −0.52, 5.77 |
| | FA | – | – | – | FA | – | – | – | FA | 2.32 | 0.04 | 0.11, 4.53 |
| | Model $R^2$ = 0.8 | | | | Model $R^2$ = 0.64 | | | | Model $R^2$ = 0.96 | | | |
| Right temporal | Volume | −0.89 | 0.02 | −1.66, −0.12 | Volume | −1.97 | 0.01 | −3.47, −0.46 | Volume | 1.51 | 0.14 | −0.47, 3.48 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.81 | | | | Model $R^2$ = 0.66 | | | | Model $R^2$ = 0.97 | | | |
| Left temporal | Volume | −0.56 | 0.20 | −1.42, 0.3 | Volume | −0.93 | 0.24 | −2.5, 0.64 | Volume | 1.85 | 0.07 | −0.17, 3.86 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.84 | | | | Model $R^2$ = 0.63 | | | | Model $R^2$ = 0.97 | | | |
| **SvPPA** | | | | | | | | | | | | |
| Right frontal | Volume | 0.69 | 0.47 | −1.2, 2.59 | Volume | 1.27 | 0.32 | −1.24, 3.78 | Volume | 0.84 | 0.58 | −2.11, 3.79 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.69 | | | | Model $R^2$ = 0.87 | | | | Model $R^2$ = 0.79 | | | |
| Left frontal | Volume | −0.46 | 0.63 | −2.34, 1.41 | Volume | 0.68 | 0.58 | −1.72, 3.08 | Volume | 1.94 | 0.11 | −0.46, 4.34 |
| | ASL | 0.38 | 0.58 | −0.97, 1.74 | ASL | 0.37 | 0.63 | −1.14, 1.87 | ASL | – | – | – |
| | FA | −0.73 | 0.39 | −2.39, 0.93 | FA | −0.81 | 0.44 | −2.88, 1.25 | FA | – | – | – |
| | Model $R^2$ = 0.64 | | | | Model $R^2$ = 0.85 | | | | Model $R^2$ = 0.83 | | | |
| Right temporal | Volume | 1.83 | 0.03 | 0.16, 3.51 | Volume | 1.61 | 0.21 | −0.89, 4.1 | Volume | −2.54 | 0.11 | −5.67, 0.59 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.71 | | | | Model $R^2$ = 0.87 | | | | Model $R^2$ = 0.81 | | | |
| Left temporal | Volume | −2.24 | 0.02 | −4.15, −0.33 | Volume | −1.73 | 0.20 | −4.35, 0.9 | Volume | 0.19 | 0.91 | −3.27, 3.66 |
| | ASL | – | – | – | ASL | – | – | – | ASL | – | – | – |
| | FA | – | – | – | FA | – | – | – | FA | – | – | – |
| | Model $R^2$ = 0.69 | | | | Model $R^2$ = 0.88 | | | | Model $R^2$ = 0.79 | | | |
| **NfvPPA** | | | | | | | | | | | | |
| Right frontal | Volume | −1.55 | 0.07 | −3.23, 0.13 | Volume | −5.28 | 0.12 | −11.89, 1.34 | Volume | −0.45 | 0.82 | −4.46, 3.55 |
| | ASL | −1.33 | 0.04 | −2.61, −0.04 | ASL | −4.39 | 0.03 | −8.35, −0.42 | ASL | – | – | – |
| | FA | −1.03 | 0.10 | −2.26, 0.21 | FA | −1.63 | 0.33 | −4.9, 1.65 | FA | – | – | – |
| | Model $R^2$ = 0.63 | | | | Model $R^2$ = 0.77 | | | | Model $R^2$ = 0.9 | | | |
| Left frontal | Volume | −0.90 | 0.25 | −2.45, 0.64 | Volume | −0.55 | 0.81 | −4.99, 3.88 | Volume | −1.25 | 0.46 | −4.61, 2.1 |
| | ASL | −0.97 | 0.08 | −2.06, 0.11 | ASL | −4.48 | 0.00 | −6.58, −2.37 | ASL | 2.33 | 0.05 | −0.05, 4.71 |
| | FA | −1.06 | 0.09 | −2.27, 0.16 | FA | −2.80 | 0.02 | −5.14, −0.47 | FA | 3.23 | 0.01 | 0.88, 5.59 |
| | Model $R^2$ = 0.72 | | | | Model $R^2$ = 0.8 | | | | Model $R^2$ = 0.9 | | | |
| Right temporal | Volume | −0.47 | 0.46 | −1.73, 0.78 | Volume | 2.56 | 0.21 | −1.45, 6.57 | Volume | −0.96 | 0.48 | −3.67, 1.74 |
| | ASL | – | – | – | ASL | −6.48 | 0.03 | −12.45, −0.52 | ASL | – | – | – |
| | FA | – | – | – | FA | −4.60 | 0.14 | −10.71, 1.51 | FA | – | – | – |
| | Model $R^2$ = 0.82 | | | | Model $R^2$ = 0.69 | | | | Model $R^2$ = 0.93 | | | |
| Left temporal | Volume | −0.43 | 0.49 | −1.65, 0.78 | Volume | 2.51 | 0.08 | −0.3, 5.32 | Volume | −1.41 | 0.39 | −4.62, 1.8 |
| | ASL | −1.02 | 0.10 | −2.23, 0.2 | ASL | −5.18 | 0.00 | −7.48, −2.88 | ASL | 1.32 | 0.40 | −1.75, 4.39 |
| | FA | −1.79 | 0.03 | −3.41, −0.17 | FA | −5.12 | 0.00 | −8.22, −2.03 | FA | 1.31 | 0.50 | −2.47, 5.1 |
| | Model $R^2$ = 0.77 | | | | Model $R^2$ = 0.78 | | | | Model $R^2$ = 0.92 | | | |

A model with only volume as an imaging predictor was compared to a model with all three imaging modalities. If the likelihood ratio test was $<0.05$, suggesting that the model with all three imaging predictors fit better, parameters for this model are shown. Otherwise, the parameters for the volume model are presented. Imaging predictors were standardized; each parameter indicates the units of raw change in the outcome per standard deviation decline in the imaging metric *b* = unstandardized coefficient.
CI = confidence interval; FA = fractional anisotropy.

tailored to each variant could yield even better measures of change (Rogalski *et al.*, 2014; Ard *et al.*, 2015; Binney *et al.*, 2017), and this should be explored using methods that combine multiple imaging modalities. In this sample, we chose to model time effects linearly to avoid the risk of over-interpreting the results. However, we note that future studies, with longer duration data or/and more frequent data acquisition, might want to examine the nature of potential non-linear trajectories. In an effort to maximize the number of subjects in this analysis, we included individuals who had been seen at irregular intervals and individuals who had not undergone all of the procedures. Thus, the studies did not

use true clinical trial methodology, and clinical features that influence enrolment criteria or dropout might differ in clinical trials and thus be associated with different effect sizes for longitudinal change. Furthermore, because we used a standard two time point approach to our power calculations, the subgroups used for sample size calculations were smaller than those used for LME analyses, and varied depending on available data. Accordingly, direct comparison of sample sizes across measures, while informative and in broad agreement with previously published data, should be interpreted with caution as is evidenced by the sometimes very wide 95% confidence intervals. Although bootstrapping can help understand potential over-fitting, the large number of measures explored in this study could result in a small sample size estimate by chance alone. Nonetheless, we recommend future publications that provide estimated sample sizes quantify uncertainty in their estimates. Despite these limitations, this is the largest, most comprehensive longitudinal dataset available in FTD, and it can provide valuable information for planning clinical trials. These data are now publicly available through LONI (https://ida.loni.usc.edu) for use by other investigators.

## Funding

This project was supported by the National Institutes of Health grant numbers AG045333, AG045390, AG032306 (H.J.R.), and AG019724 and AG023501 (B.L.M.) and the Hillblom Network.

## Competing interests

Outside of the submitted work, A.L.B reports grants from BMS, Biogen, C2N, Tau Consortium, Bluefield Project to Cure FTD, CBD Solutions, Genentech, TauRx, Forum, NIH, and personal fees from Abbvie, Asceneuron, Celgene, Roche, Novartis, Merck, and non-financial support from Quanterix. B.F.B reports grants from GE Healthcare, NIH, Mangurian Foundation, FORUM Pharmaceuticals, C2N Diagnostics, Axovant, and personal fees from Scientific Advisory Board - Tau Consortium. D.S.K. reports grants from Biogen and Lilly A4 study, and personal fees from Lundbeck Pharma and the DIAN study. H.J.R. reports grants from the NIH and California Department of Public Health during the conduct of the study, as well as grants from Quest Diagnostics and Biogen Pharmaceuticals, outside the submitted work. B.L.M., MD receives grant support from the NIH/NIA and the Centers for Medicare & Medicaid Services (CMS) as grants for the Memory and Aging Center. As an additional disclosure, B.L.M. serves as Medical Director for the John Douglas French Foundation; Scientific Director for the Tau Consortium; Director/Medical Advisory Board of the Larry L. Hillblom Foundation; Scientific Advisory Board Member for the National Institute for Health Research Cambridge Biomedical Research Centre and its subunit, the Biomedical Research Unit in Dementia (UK); and Board Member for the American Brain Foundation (ABF). AMS receives support from the Larry L. Hillblom Foundation and the NIH.

## Supplementary material

Supplementary material is available at *Brain* online.

## References

Ard MC, Raghavan N, Edland SD. Optimal composite scores for longitudinal clinical trials under the linear mixed effects model. Pharm Stat 2015; 14: 418–26.

Binney RJ, Pankov A, Marx G, He X, Mckenna F, Staffaroni AM, et al. Data-driven regions of interest for longitudinal change in three variants of frontotemporal lobar degeneration. Brain Behav 2017; 7(4): e00675.

Boxer AL, Gold M, Huey E, Hu WT, Rosen H, Kramer J, et al. The advantages of frontotemporal degeneration drug development (part 2 of frontotemporal degeneration: the next therapeutic frontier). Alzheimer's Dement 2013; 9(2): 189–98.

Boxer AL, Knopman DS, Kaufer DI, Grossman M, Onyike C, Graf-Radford N, et al. Memantine in patients with frontotemporal lobar degeneration: a multicentre, randomised, double-blind, placebo-controlled trial. Lancet Neurol 2013; 12: 149–56.

Busner J, Targum SD. The clinical global impressions scale: applying a research tool in clinical practice. Psychiatry (Edgmont) 2007; 4: 28–37.

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.

Daly E, Zaitchik D, Copeland M, Schmahmann J, Gunther J, Albert M. Predicting conversion to Alzheimer disease using standardized clinical information. Arch Neurol 2000; 57: 675–80.

Delis DC, Kaplan E, Kramer JH. Delis-Kaplan Executive Function System (DKEFS): Examiner's manual. San Antonio, TX: The Psychological Corporation; 2001.

Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test - second edition. Adult version. Manual. San Antonio, TX: Psychological Corporation; 2000.

Dopper EGP, Chalos V, Ghariq E, den Heijer T, Hafkemeijer A, Jiskoot LC, et al. Cerebral blood flow in presymptomatic MAPT and GRN mutation carriers: a longitudinal arterial spin labeling study. NeuroImage Clin 2016; 12: 460–5.

Du AT, Jahng GH, Hayasaka S, Kramer JH, Rosen HJ, Gorno-Tempini ML, et al. Hypoperfusion in frontotemporal dementia and Alzheimer disease by arterial spin labeling MRI. Neurology 2006; 67: 1215–20.

Dunn LM, Dunn LM. Peabody Picture Vocabulary Test - Revised. Circle Pines, MN: American Guidance Service; 1981.

Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. Stat Med 2012; 31: 2973–84.

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.

Gordon E, Rohrer JD, Kim ML, Omar R, Rossor MNM, Fox FN, et al. Measuring disease progression in frontotemporal lobar degeneration: a clinical and MRI study. Neurology 2010; 74: 666–73.

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.

Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. Neurology 2011; 76: 1006–14.

Howard D, Patterson K. Pyramids and palm trees: A test of semantic access from pictures and words. Bury St. Edmunds, UK: Thames Valley Test Company; 1992.

Hyman DM, Piha-Paul SA, Won H, Rodon J, Saura C, Shapiro GI, et al. HER kinase inhibition in patients with HER2- and HER3-mutant cancers. Nature 2018; 554: 189–94.

Hyman DM, Taylor BS, Baselga J. Implementing genome-driven oncology. Cell 2017; 168: 584–99.

Ittner LM, Götz J. Amyloid-β and tau — a toxic pas de deux in Alzheimer's disease. Nat Rev Neurosci 2011; 12: 67–72.

Kaplan E, Goodglass H, Weintraub S. Boston naming test. Philadelphia: Lea & Febiger; 1983.

Katz R. Biomarkers and surrogate markers: an FDA perspective. NeuroRx 2004; 1: 189–95.

Khan BK, Yokoyama JS, Takada LT, Sha SJ, Rutherford NJ, Fong JC, et al. Atypical, slowly progressive behavioural variant frontotemporal dementia associated with C9ORF72 hexanucleotide expansion. J Neurol Neurosurg Psychiatry 2012; 83: 358–64.

Knopman DS, Jack CR, Kramer JH, Boeve BF, Caselli RJ, Graff-Radford NR, et al. Brain and ventricular volumetric changes in frontotemporal lobar degeneration over 1 year. Neurology 2009; 72: 1843–9.

Knopman DS, Kramer JH, Boeve BF, Caselli RJ, Graff-Radford NR, Mendez MF, et al. Development of methodology for conducting clinical trials in frontotemporal lobar degeneration. Brain 2008; 131: 2957–68.

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.

Lam BYK, Halliday GM, Irish M, Hodges JR, Piguet O. Longitudinal white matter changes in frontotemporal dementia subtypes. Hum Brain Mapp 2014; 35: 3547–57.

Lee D-H, Park JW, Park S-H, Hong C, Dubovsky E. Have you ever seen the impact of crossing fiber in DTI?: demonstration of the corticospinal tract pathway. PLoS One 2015; 10: e0112045.

Lee SE, Sias AC, Mandelli ML, Brown JA, Brown AB, Khazenzon AM, et al. Network degeneration and dysfunction in presymptomatic C9ORF72 expansion carriers. Neuroimage Clin 2017; 14: 286–97.

Mahoney CJ, Simpson IJA, Nicholas JM, Fletcher PD, Downey LE, Golden HL, et al. Longitudinal diffusion tensor imaging in frontotemporal dementia. Ann Neurol 2015; 77: 33–46.

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.

Pankov A, Binney RJ, Staffaroni AM, Kornak J, Attygalle S, Schuff N, et al. Data-driven regions of interest for longitudinal change in frontotemporal lobar degeneration. Neuroimage Clin 2016; 12: 332–40.

Panza F, Seripa D, Solfrizzi V, Imbimbo BP, Lozupone M, Leo A, et al. Emerging drugs to reduce abnormal β-amyloid protein in Alzheimer's disease patients. Expert Opin Emerg Drugs 2016; 21: 377–91.

Perry D, Miller B. Frontotemporal dementia. Semin Neurol 2013; 33: 336–41.

Perry DC, Rosen HJ. Frontotemporal dementia. In: Geschwind MD, Racine Belkoura C, editors. Non-Alzheimer's and atypical dementia. John Wiley & Sons, Ltd.; 2016. p. 49–63.

Pfeffer RI, Kurosaki TT, Harrah CH, Chance JM, Filos S. Measurement of functional activities in older adults in the community. J Gerontol 1982; 37: 323–9.

Possin KL, Chester SK, Laluz V, Bostrom A, Rosen HJ, Miller BL, et al. The frontal-anatomic specificity of design fluency repetitions and their diagnostic relevance for behavioral variant frontotemporal dementia. J Int Neuropsychol Soc 2012; 18: 834–44.

Ramanan S, Bertoux M, Flanagan E, Irish M, Piguet O, Hodges JR, et al. Longitudinal executive function and episodic memory profiles in behavioral-variant frontotemporal dementia and Alzheimer's disease. J Int Neuropsychol Soc 2017; 22: 1–10.

Ranasinghe KG, Rankin KP, Pressman PS, Perry DC, Lobach IV, Seeley WW, et al. Distinct subtypes of behavioral variant frontotemporal dementia based on patterns of network degeneration. JAMA Neurol 2016; 73: 1078.

Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 2011; 134: 2456–77.

Rogalski E, Cobia D, Martersteck A, Rademaker A, Wieneke C, Weintraub S, et al. Asymmetry of cortical decline in subtypes of primary progressive aphasia. Neurology 2014; 83: 1184–91.

Rohrer JD, Nicholas JM, Cash DM, van Swieten J, Dopper E, Jiskoot L, et al. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis. Lancet Neurol 2015; 14: 253–62.

Sakpal TV. Sample size estimation in clinical trial. Perspect Clin Res 2010; 1: 67–9.

Santillo AF, Martensson J, Lindberg O, Nilsson M, Manzouri A, Landqvist Waldo M, et al. Diffusion tensor tractography versus volumetric imaging in the diagnosis of behavioral variant frontotemporal dementia. PLoS One 2013; 8: e66932.

Schubert S, Leyton CE, Hodges JR, Piguet O. Longitudinal memory profiles in behavioral-variant frontotemporal dementia and Alzheimer's disease. J Alzheimer's Dis 2016; 51: 775–82.

Schwab R, England A. Projection technique for evaluating surgery in Parkinson's disease. 1969.

Seeley WW. Mapping neurodegenerative disease onset and progression. Cold Spring Harb Perspect Biol 2017; 9: a023622.

Spinelli E, Mandelli ML, Santos M, Vinceti G, Watson C, Agosta F, et al. In vivo correlates of pathological diagnosis in primary progressive aphasia (P4.009). Neurology 2016; 86: P4.009.

Staffaroni AM, Elahi FM, McDermott D, Marton K, Karageorgiou E, Sacco S, et al. Neuroimaging in dementia. Semin Neurol 2017; 37: 510–37.

Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol 1935; 18: 643–62.

Tosun D, Schuff N, Rabinovici GD, Ayakta N, Miller BL, Jagust W, et al. Diagnostic utility of ASL-MRI and FDG-PET in the behavioral variant of FTD and AD. Ann Clin Transl Neurol 2016; 3: 740–51.

Tsai RM, Boxer AL. Therapy and clinical trials in frontotemporal dementia: past, present, and future. J Neurochem 2016; 138: 211–21.

Tu S, Leyton CE, Hodges JR, Piguet O, Hornberger M. Divergent longitudinal propagation of white matter degradation in logopenic and semantic variants of primary progressive aphasia. J Alzheimer's Dis 2015; 49: 853–61.

Vandenberghe R, Rinne JO, Boada M, Katayama S, Scheltens P, Vellas B, et al. Bapineuzumab for mild to moderate Alzheimer's disease in two global, randomized, phase 3 trials. Alzheimers Res Ther 2016; 8: 18.

Wechsler D. Wechsler Adult Intelligence Scale. San Antonio, TX: Psychological Corporation; 1997.

Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 1982; 17: 37–49.

Zhang Y, Tartaglia MC, Schuff N, Chiang GC, Ching C, Rosen HJ, et al. MRI signatures of brain macrostructural atrophy and microstructural degradation in frontotemporal lobar degeneration subtypes. J Alzheimers Dis 2013; 33: 431–44.

Zu C, Jie B, Liu M, Chen S, Shen D, Zhang D, et al. Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. Brain Imaging Behav 2016; 10: 1148–59.