Article

Short-interval observational data to inform clinical trial design in Huntington's Disease

Nicola Z. Hobbs PhD[1], Ruth E. Farmer MSc[2], Elin M. Rees MSc[1], James H. Cole PhD[1], Salman Haider MD[1], Ian B Malone PhD[3], Reiner Sprengelmeyer PhD[4], Hans Johnson PhD[5], Hans-Peter Mueller PhD[4], Sigurd D. Sussmuth MD[4], Raymund A.C. Roos MD[6], Alexandra Durr PhD[7], Chris Frost MA[2], Rachael I. Scahill PhD[1], Bernhard Landwehrmeyer MD[4], Sarah J Tabrizi PhD[1]

Affiliations

(1) Department of Neurodegenerative Disease, UCL Institute of Neurology, University College London, Queen Square, London, WC1N 3BG, UK

(2) Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

(3) Dementia Research Centre, UCL Institute of Neurology, University College London, Queen Square, London, WC1N 3BG, UK

(4) Department of Neurology, Ulm University, Ulm, Germany

(5) Department of Psychiatry, University of Iowa, Iowa City, Iowa, USA

(6) Department of Neurology, Leiden University Medical Centre, Leiden, Netherlands

(7) ICM (Brain and Spine Institute), APHP- Department of Genetics and INSERM UMR S679, Salpêtrière University Hospital, Paris, France

Corresponding author contact information:

Prof. Sarah Tabrizi: s.tabrizi@prion.ucl.ac.uk

Address: Box 104, Queen Square, London, WC1N 3BG, UK.

Tel: 00 44 203 448 4053. Fax: 00 44 207 611 0129.

Word count = 3397;

Search Terms : MRI, Huntington's Disease, Observational study.

ABSTRACT

Objectives: To evaluate candidate outcomes for disease-modifying trials in Huntington's Disease (HD) over 6-, 9- and 15-month intervals, across multiple domains. To present guidelines on rapid efficacy readouts for disease-modifying trials.

Methods: 40 controls and 61 HD patients, recruited from four EU sites, underwent 3T MRI and standard clinical and cognitive assessments at baseline, 6- and 15-months. Neuroimaging analysis included global and regional change in macrostructure (atrophy and cortical thinning) and microstructure (diffusion metrics). The main outcome was longitudinal Effect Size (ES) for each outcome. Such ES can be used to calculate sample-size requirements for clinical trials for hypothesised treatment efficacies.

Results: Longitudinal changes in macrostructural neuroimaging measures such as caudate atrophy and ventricular expansion were significantly larger in HD than controls, giving rise to consistently large ES over the 6-, 9- and 15-month intervals. Analogous ES for cortical metrics were smaller with wide confidence intervals. Microstructural (diffusion) neuroimaging metrics ES were also typically smaller over the shorter intervals, although caudate diffusivity metrics performed strongly over 9- and 15-months. Clinical and cognitive outcomes exhibited small longitudinal ESs, particularly over 6- and 9-month intervals, with wide confidence intervals, indicating a lack of precision.

Conclusions: To exploit the potential power of specific neuroimaging measures such as

caudate atrophy in disease-modifying trials, we propose their use as (1) initial short-term readouts in early phase/proof-of-concept studies over six or nine months, and (2) secondary end-points in efficacy studies over longer periods such as 15 months.

INTRODUCTION

Major efforts are being invested in the development of disease-modifying therapies for neurodegenerative disorders such as Huntington's disease (HD).[1] Testing their efficacy in clinical trials is a long and expensive process, with low success rates compared with other branches of medicine.[2] In HD, no phase III studies of putative disease-modifying treatments have been successful, despite many showing promise during early testing.

A wealth of observational data suggests that biomarkers of disease progression may facilitate the evaluation of disease-modifying therapies.[3-6] MRI-derived neuroimaging measures appear particularly powerful, with data suggesting that substantially fewer patients would be required to detect a reduction in rate of change in MRI biomarkers, compared with clinical measures.[3-9] However, many biomarkers have only been evaluated over intervals ≥12 months.

It may be advantageous for clinical trials to have efficacy readouts over short intervals such as six months, especially during the early phases, in order to provide confidence-instilling data that the trial should progress to a larger scale. However, the use of short-interval biomarkers in clinical trials is critically dependent on their validation in longitudinal observational studies over the same time frame.

Our objectives were to evaluate candidate outcomes for HD trials over 6-, 9- and 15-month intervals, across neuroimaging, clinical and cognitive domains. Based on our findings, we present guidelines on the selection of outcomes for rapid readouts in clinical trials. It is hoped these data will directly inform the design of HD trials, facilitating the evaluation of treatments designed to slow the course of this devastating disease.

METHODS

Study Design

This was a longitudinal, case-control observational study in HD. Assessments were performed at baseline, six and 15 months. The study was approved by the local ethical committees.

Participants

Between March and October 2011, 40 controls and 61 HD patients were enrolled into Work Package 2 of the PADDINGTON study [10] at Leiden (Netherlands), London (UK), Paris (France) and Ulm (Germany). Patients were recruited from research centres. Controls were spouses, partners or gene-negative siblings in order to match patients to controls as closely as possible in terms of age, education level, background and home life. Patients were ideally required to be at stage I of the disease,[10] defined by a Unified Huntington's Disease Rating Scale (UHDRS)[11] Total Functional Capacity (TFC) ≥ 11, indicating good capacity in functional realms; however, five patients were granted waivers for not fulfilling this TFC criteria, as described in the Results. Inclusion criteria included participants being 18–65 years of age, free from major psychiatric and

concomitant neurological disorders, not currently participating in a clinical trial and able to tolerate and safely undergo MRI. Written informed consent was obtained from each subject.

Procedures

Clinical features were assessed using the UHDRS version '99. This included the Total Motor Score (TMS) which measures a range of motor features characteristically impaired in HD including gait, tongue protrusion, ocular function and postural stability; and the TFC scale which measures five components of daily living, including the capacity to work, manage finances and carry out domestic chores. The clinical examination was performed by raters certified by the European Huntington's Disease Network (EHDN) UHDRS-TMS online certification (www.euro-hd.net).

Cognitive features were assessed using the core EHDN cognitive battery, which consists of standard pencil and paper clinical neuropsychological tasks. All raters were trained on the battery and all tests were scripted. Each task is described in the Supplemental Methods.

MRI acquisition

3T MRI (T1-, T2- and diffusion-weighted) were acquired based on protocols standardised for multi-site use.[6,10,12] Scan acquisition protocols have been described previously.[10] Quality control was performed on all datasets in pseudo-real time and rescans were requested where necessary. Data were pseudoanonymised and archived

on a secure webportal. To avoid potential bias all image analysis was performed blinded to groupings.

MRI: macrostructural (volumetric) analysis

Pre-defined regions-of-interest (ROIs) for the volumetric analysis included the caudate, putamen, white-matter, grey-matter, whole-brain, lateral ventricles and corpus callosum. Cortical thinning was also examined over each lobe (parietal, occipital, temporal and frontal).

The software package MIDAS[13] was used to delineate the whole-brain, caudate, corpus callosum and ventricles at baseline.[10] Change in whole-brain, caudate and ventricular volume over the scanning interval was estimated using the Boundary Shift Integral (BSI) technique,[14] optimised for multi-site data,[15] within MIDAS software. The BSI is a semi-automated tool which measures volume change over time (atrophy) directly from within-subject registered scan pairs. Change in corpus callosum and putamen volume was estimated by delineating the structures at both time-points, either manually[11] (for all corpus callosum measurements) or with BRAINS3 software[6,16] (for all putamen measurements) and subtracting the volumes at each time-point. Grey-matter and white-matter volume changes were computed using a fluid-registration approach.[5,17,18]

Cortical thickness measures were computed using FreeSurfer software (http://surfer.nmr.mgh.harvard.edu/; version 5.3.0). All scans were run through the longitudinal pipeline[19] and thickness estimates (mm) were extracted from each region defined by the Desikan-Killiany Atlas and averaged within lobes.[20]

Full details of all volumetric image analysis are provided in the Supplemental Methods.

MRI: microstructural (diffusion) analysis

Diffusion metrics of fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD) and radial diffusivity (RD) were generated over pre-defined ROIs (white-matter, corpus callosum, caudate and putamen) for all three visits using a longitudinal registration pipeline. In brief, a common ROI mask was defined in a temporally unbiased 'mid-space' based on within-subject registration of T1 images, before being non-linearly registered to each individual's native FA images for each visit. The mean values were then calculated across all included voxels for the four DTI metrics. This analysis is described in detail in the Supplemental Methods.

All segmentations and registrations were visually inspected for accuracy by trained analysts, blinded to diagnosis. Excluded data-points are described in Supplemental End-point Quality Control data.

Statistical analysis

Statistical analysis was performed by an independent team according to a predefined analysis plan. The repeated measures of each outcome variable were analysed using generalised least squares regression models, with variances of the outcome (and correlations between pairs of measures) allowed to differ both by group and by visit. The models included a group factor (HD or Control), calendar time from baseline (in days) and a quadratic term to allow non-linear change over the three visits to be modelled. The use of GLS models that jointly model all available outcomes provides

some additional protection against the impact of missing values. Data only requires a "missing at random" assumption rather than the more restrictive "missing completely at random" assumption to give unbiased estimates.[21] Where outcomes directly measured changes (such as whole-brain atrophy between two visits) the outcome variables in the statistical models were change between baseline and six months (i.e. 6-month interval), change between baseline and 15 months (i.e. 15-month interval) and change between six and 15 months (i.e. 9-month interval). Otherwise outcomes were measures made at baseline, six and 15 months. Linear and quadratic effects of time were included in all models with estimated between-group differences for the 6-, 9- and 15-month intervals calculated using appropriate linear combinations of model parameters. All analyses adjusted for baseline age, gender and study site as well as interactions with the linear and quadratic effects of time. This was due to an a priori belief that age, gender and study site might affect slopes (and rates of change in slopes) as well as absolute levels of the outcomes. Models for non-imaging outcomes adjusted additionally for educational level (an ordered categorical variable treated as a continuous covariate) and its interactions with linear and quadratic effects of time because education level may affect performance on such outcomes, and education levels were expected to differ systematically between HD and controls.

Longitudinal Effect Sizes (ES) with 95% confidence intervals (CIs) for the difference in change over each interval were calculated as the covariate-adjusted difference in the mean of the change between HD participants and controls, divided by the estimated residual standard deviation (SD) of change in HD participants. Expression of results as

(unit-free) ES permits comparison of changes measured using different metrics. The square of ES is inversely related to sample-size requirements for clinical trials under the assumption that a 100% effective treatment will reduce the mean rate of change in HD cases to that in healthy controls without affecting the variability in these rates.[22] 95% CIs for the ES were calculated using bias corrected and accelerated (BCa) bootstrapping, with 2000 replications [23]. Here an ES of two implies that the mean change in HD is two SD away from that in controls. No formal criteria was used to assess "size" of ES. Since thresholds for such criteria could be argued to be arbitrary, the approach taken was to consider ES in relation to each other at each time point, and to evaluate whether the estimated ES and 95% confidence intervals translated into feasible sample size estimates for the specific context of HD clinical trials. No adjustment for multiple comparisons was made since there is independent scientific interest in each of the variables.[24] Throughout, a cut-off of p=0.05 was used to establish formal statistical significance, with the actual p-values also considered in the interpretation of results. All analysis was performed in STATA v12.

RESULTS

Participants

At baseline, five HD participants were granted waivers for being outside disease-stage 1; four were stage 2, one was stage 3.[10] All controls and 59/61 HD participants returned for the 6-month assessment; HD non-attendance was due to illness (n=2), both returned for the 15-month visit. 37/40 controls and 56/61 HD participants returned for the 15-month assessment; HD drop-out was due to disease-related burden (n=1), inability to

tolerate scanning (n=1), treatment for cancer (n=1) and psychiatric burden resulting in the site investigator withdrawing the participant (n=2). Drop out in the control group was due to being the spouse of a withdrawn HD participant (n=1) or personal issues unrelated to the study (n=2).

Table 1: Participant demographics at baseline

Characteristic	**Controls (N=40)**		**HD Stage I (N=61)**	
Age (Years) Mean (SD) Min – Max	51.4 (8.4)	29.0 – 66.6	48.7 (10.8)	23.5 – 7.3
Gender				
Female N (%)	23	(57.5%)	37	(60.7%)
Site				
Leiden N (%)	10	(25%)	17	(27.87%)
London N (%)	10	(25%)	16	(26.23%)
Paris N (%)	10	(25%)	13	(21.31%)
Ulm N (%)	10	(25%)	15	(24.59%)
TMS Mean (SD) Min - Max	1.4 (1.9)	0-7	20.1 (10.7)	6-58
TFC Mean (SD) Min - Max	13.0 (0.2)	12-13	11.7 (1.5)	5-13
CAG Mean (SD) Min - Max			43.8 (3.2)	39 – 54
Disease Burden Score * Mean (SD) Min - Max			376.5 (85.2)	226.4 - 59.2
TFC by site N(%)				
TFC 11-13 (HD Stage 1)			56	(91.80%)

TFC 7-10 (HD Stage 2)			4[a]	(6.56%)
TFC 3-6 (HD Stage 3)			1[b]	(1.64%)

All study participants attended for at least 1 follow up clinical visit
SDMT = Symbol Digit Modality Test; TMS = Total Motor Score; TFC = Total Functional Capacity
* Penny[25] Disease Burden Formula: Age x (CAG - 35.5)
[a] 3 London, 1 Paris.[b] Paris

Age and gender were well-balanced between groups (Table 1). Within the HD group, CAG, disease burden[25] and TFC were well-balanced between sites (Supplemental Table 1). The average intervals in months (mean (SD)) between assessments in the HD group were 5.76 (1.36), 9.12 (0.99) and 14.88 (1.33). In the control group the intervals were 5.48 (1.08), 9.08 (0.88) and 14.50 (1.09).

Effect sizes

ES for the difference in 6-, 9- and 15-month change between HD participants and controls are presented in Table 2. Unadjusted baseline, 6- and 15-month findings for each outcome, with the number of data points for each variable, are presented by group in Supplemental Tables 3 and 4, with adjusted between-group differences in change over the 6-, 9-, and 15-month intervals.

For clinical applicability, Table 2 should be viewed in conjunction with Figure 1, which depicts the relationship between ES and sample-size requirements for disease-modifying clinical trials (where the outcome is a single change measured between two time points) for varying assumed treatment efficacies.

Table 2: 6-, 9- and 15-month Effect Size Estimates

	Effect Size Estimate (95% CI)		
	6-month interval	**9-month interval**	**15-month interval**
Cognitive battery			
Letter Fluency	0.13 (-0.40, 0.60)	0.62 (-0.07, 1.18)	0.66 (-0.03, 1.32)
Category Fluency	0.23 (-0.21, 0.66)	0.13 (-0.42, 0.66)	0.35 (-0.20, 0.89)
HVLT delayed recall	0.49 (-0.01, 0.93)	0.00 (-0.53, 0.53)	0.50 (-0.12, 1.03)
HVLT total correct	0.12 (-0.36, 0.59)	0.12 (-0.33, 0.61)	0.21 (-0.18, 0.58)
HVLT Recognition	0.19 (-0.15, 0.45)	-0.26 (-0.69, 0.08)	-0.16 (-0.84, 0.32)
SDMT	0.64 (0.08, 1.15)	0.34 (-0.11, 0.81)	0.80 (0.34, 1.25)
Trail A Time (seconds)	0.21 (-0.10, 0.47)	-0.06 (-0.37, 0.31)	0.21 (-0.12, 0.57)
Trail B Time (seconds)	0.11 (-0.27, 0.44)	-0.23 (-0.68, 0.16)	-0.07 (-0.49, 0.25)
Stroop Word	0.29 (-0.09, 0.57)	0.06 (-0.26, 0.45)	0.31 (-0.08, 0.61)
Stroop Colour	0.25 (-0.19, 0.68)	0.19 (-0.23, 0.59)	0.36 (-0.03 , 0.71)
Stroop Interference	0.17 (-0.19, 0.54)	0.30 (-0.11, 0.69)	0.49 (-0.03 , 0.94)
UHDRS clinical scales			
TMS (square root)	0.05 (-0.47, 0.61)	0.58 (0.09, 1.10)	0.55 (0.08, 1.12)
TFC score	0.33 (-0.53, 1.33)	0.18 (-1.05, 1.32)	0.39 (-0.48, 1.24)
Microstructural (diffusion) neuroimaging metrics			
Caudate FA	0.37 (-0.13, 0.83)	0.29 (-0.11, 0.65)	0.52 (0.12 , 0.88)
Caudate MD (mm^2/s)	0.54 (0.20, 0.83)	0.62 (0.17, 1.03)	1.11 (0.77, 1.43)
Caudate RD (mm^2/s)	0.52 (0.18, 0.82)	0.61 (0.18, 1.02)	1.07 (0.73 , 1.39)
Caudate AD (mm^2/s)	0.56 (0.21, 0.86)	0.63 (0.15, 1.06)	1.174 (0.84, 1.49)
Putamen FA	-0.04 (-0.36, 0.30)	-0.21 (-0.56, 0.15)	-0.27 (-0.65, 0.14)
Putamen MD (mm^2/s)	0.43 (0.15, 0.72)	0.29 (-0.07, 0.64)	0.72 (0.38, 1.02)
Putamen RD (mm^2/s)	0.33 (0.06, 0.60)	0.22 (-0.15, 0.56)	0.57 (0.23, 0.87)

Putamen AD (mm^2/s)	0.55 (0.23, 0.85)	0.38 (0.03, 0.72)	0.92 (0.53, 1.26)
White Matter FA	0.23 (-0.16, 0.64)	-0.09 (-0.48, 0.29)	0.17 (-0.28, 0.65)
White Matter MD (mm^2/s)	0.50 (0.07, 0.93)	0.19 (-0.15, 0.54)	0.62 (0.20, 1.10)
White Matter RD (mm^2/s)	0.39 (-0.05, 0.79)	0.10 (-0.21, 0.42)	0.51 (0.08, 0.94)
White matter AD (mm^2/s)	0.50 (0.08, 0.89)	0.28 (-0.07, 0.81)	0.61 (0.21, 1.14)
Corpus Callosum FA	0.43 (0.11, 0.82)	0.15 (-0.21, 0.47)	0.68 (0.17, 1.15)
Corpus Callosum MD (mm^2/s)	0.25 (-0.18, 0.76)	0.15 (-0.30, 0.56)	0.30 (-0.12, 0.90)
Corpus Callosum RD (mm^2/s)	0.37 (-0.00 , 0.88)	0.10 (-0.32, 0.50)	0.41 (-0.05, 1.03)
Corpus Callosum AD (mm^2/s)	0.02 (-0.35, 0.38)	0.24 (-0.24, 0.64)	0.21 (-0.16, 0.72)
Macrostructural (volumetric) neuroimaging metrics			
Caudate atrophy, CBSI (% baseline)	0.70 (0.36, 1.02)	0.64 (0.32, 0.98)	1.19 (0.74, 1.69)
Whole-brain atrophy, BBSI (% baseline)	0.48 (0.16, 0.77)	0.70 (0.31, 1.06)	0.87 (0.47, 1.20)
Ventricular expansion, VBSI (mls)	0.79 (0.41, 1.14)	0.93 (0.55, 1.28)	1.03 (0.67, 1.32)
Grey matter atrophy (% baseline)	0.77 (0.24, 1.23)	0.61 (0.30, 1.10)	0.86 (0.55, 1.22)
White matter atrophy (% baseline)	0.62 (0.26, 1.03)	0.93 (0.57, 1.28)	0.96 (0.59, 1.33)
Putamen atrophy (% baseline)	0.10 (-0.19, 0.40)	0.54 (0.20, 0.90)	0.78 (0.33, 1.18)
Corpus callosal atrophy (% baseline)	0.11 (-0.27, 0.56)	0.17 (-0.21, 0.61)	0.21 (-0.19, 0.63)
Macrostructural (cortical thinning) neuroimaging metrics			
Frontal lobe cortical thinning (mm)	-0.10 (-0.52, 0.29)	-0.06 (-0.51, 0.42)	-0.17 (-0.76, 0.41)
Parietal lobe cortical thinning (mm)	0.04 (-0.32, 0.42)	0.25 (-0.15, 0.65)	0.38 (-0.11, 0.86)
Temporal lobe cortical thinning (mm)	0.29 (-0.15, 0.75)	0.06 (-0.323, 0.51)	0.25 (-0.12, 0.70)
Occipital lobe cortical thinning (mm)	0.30 (-0.16, 0.77)	0.22 (-0.20, 0.67)	0.51 (0.01, 1.00)

ES estimates and 95% bias corrected and accelerated CIs over 6-, 9- and 15-month intervals for differences between change in HD and control participants. All analyses adjusted for age, gender and study site as well as interactions with the linear and quadratic effects of time. Models for non-imaging outcomes adjusted additionally for educational level and its interactions with linear and quadratic effects of time. Expression of results as ES permits comparison of changes measured using different metrics. Such ES (when squared) are inversely related to sample-size requirements for clinical trials under the assumption that a 100% effective treatment will reduce the mean rate of change in HD to that in healthy controls, without affecting the variability.

Macrostructural neuroimaging measures

Longitudinal atrophy of the caudate, white-matter, grey-matter and whole-brain, and expansion of the lateral ventricles, produced relatively large ES over 6-, 9- and 15-month intervals (Table 2); with all between-group differences statistically significant ($p<0.05$, Supplemental Table 4). ES for these metrics were relatively consistent in that they tended to change in magnitude relative to the interval size. Caudate atrophy and ventricular expansion performed particularly strongly over the 6-month interval.

Putamen atrophy ES were small and not statistically significant over the 6-month interval (ES 0.101; 95% CI -0.187, 0.397) but performed more strongly over 9- and 15-months, although ES were smaller than for the caudate and the other more global atrophy metrics listed above (Table 2).

Corpus callosal atrophy was not significantly higher in patients than controls for all time intervals examined (Supplemental Table 3).

Cortical thinning ES were small and between-group differences were only statistically significant for the occipital cortex over the 15-month interval ($p=0.032$, Supplemental Table 3); however this ES was relatively small with a wide CI (0.512; 95% CI 0.011, 0.997).

Microstructural neuroimaging measures

The microstructural (diffusion) metrics had typically smaller ES than the macrostructural atrophy measurements, although the caudate diffusivity metrics performed strongly

(Table 2, Supplemental Table 3). In particular, caudate MD produced ES comparable to caudate atrophy over the 9- and 15-month intervals.

FA ES were small and there was little evidence of statistically significant between-group differences for all structures examined (caudate, putamen, global white-matter and corpus callosum), particularly over short intervals (Supplemental Table 3).

Clinical measures

The standard clinical scales examined (TFC and TMS) performed relatively poorly. Between-group differences in TFC were not statistically significant over 6-, 9- or 15-month intervals (Supplemental Table 3) and corresponding ES were small, with CIs spanning zero. TMS performed more strongly than TFC over the 9- and 15-month intervals, with significant between-group differences and larger ES, although the CIs surrounding the ES estimates were wide (TMS over 15 months; ES 0.545 (95% CI: 0.075, 1.123)).

Cognitive measures

Changes in the majority of tasks in the cognitive battery did not differ significantly between HD and controls over all intervals examined (Table 2, Supplemental Table 3). The Symbol Digit Modality Task (SDMT) was the most promising non-imaging measure with an ES of 0.799 (95% CI: 0.344 to 1.254) over 15 months.

DISCUSSION

Employing a multi-site study design with variable, short-interval observational periods, we report 6-, 9- and 15-month ES for a range of candidate biomarker outcomes for HD trials across multiple assessment modalities (macro- and micro-structural neuroimaging, clinical and cognitive). Reported ES can be used with a standard formula to calculate sample-size requirements for disease-modifying clinical trials[22] (Figure 1). This is the first time that ES have been reported over the short intervals of six and nine months. It is hoped that these data will be used to directly inform disease-modifying clinical trial design.

Key Results

Longitudinal changes in macrostructural neuroimaging measures such as caudate atrophy and ventricular expansion in early HD subjects were larger than those in controls giving rise to consistently large ES over the 6-, 9- and 15-month intervals, in agreement with previous multi-site observational findings over periods of 12-months and longer.[4,5,7] Analogous ES for cortical metrics were smaller, particularly over the shorter intervals. Although cortical thinning was recently used as an outcome measure in the PRECREST trial over a 6-month interval[26], our findings suggest it has limited longitudinal sensitivity and would require substantially larger sample sizes than the other macrostructural metrics reported here. Microstructural (diffusion) neuroimaging metrics ES were also typically smaller over the shorter intervals, although caudate diffusivity metrics performed strongly over 9- and 15-months, in line with the most promising atrophy measures. To our knowledge, this is the first longitudinal multi-site

study to examine change in diffusion metrics in HD. Findings are encouraging, particularly within the striatal grey matter, in accordance with a recent report over 18 months in a single-site study.[3]

Clinical and cognitive outcomes exhibited small longitudinal ESs, particularly over 6- and 9-month intervals, with wide confidence intervals, indicating a lack of precision. Of note, SDMT appeared particularly promising over the 6-month interval, producing ES comparable with caudate atrophy, although with noticeably wider confidence intervals. However, this result was not replicated over the 9-month interval, suggesting it to be a chance finding. Over 15 months, SDMT performed strongly, producing ES comparable with putamen atrophy. These longer-interval findings are in line with previous reports over 12- and 24-months, showing SDMT to be one of the most promising cognitive outcomes.[4,5,8,27]

Interpretation: Clinical application

To interpret findings within the context of designing disease-modifying clinical trials in HD, we must consider that although certain neuroimaging measures appear to be particularly powerful, they would not be accepted as primary end-points in trials since they do not provide a direct measure of how the patient feels, functions or survives (www.fda.gov). Hence, to exploit the potential of these neuroimaging measures, we propose their use as: (1) initial short-term readouts in early phase/proof-of-concept (PoC) studies over six or nine months; (2) interim or safety readouts over six or nine

months in longer, larger efficacy studies (e.g. Phase III), and as; (3) secondary end-points in efficacy studies over longer periods such as 15 months.

Short-term readouts

Macrostructural neuroimaging measures such as caudate atrophy and ventricular expansion may be able provide early confidence-instilling readouts in Phase II PoC studies over intervals such as 6- and 9- months, where the goal would be to assure safety and gather initial evidence that the therapy had promising properties. Encouraging findings from such readouts would facilitate the decision whether to further invest in the therapy, increasing participant numbers and trial duration. An adaptive approach such as this based on early, meaningful data could improve the viability of disease-modifying clinical trials in HD.

Interim read-outs and secondary end-points

Once sufficiently powered, disease-modification could be demonstrated in large-scale Phase II/III efficacy studies of longer duration such as 15 months, using approved clinical measures such as TMS as the primary end-point, and specific neuroimaging metrics as secondary end-points. Supportive data from a strong neuroimaging biomarker programme would be important in demonstrating disease modification.

Figure 2 provides an example of how the ES data presented in Table 2 could be used to inform clinical trial design. Sample-size requirements are presented for the most promising outcomes from each assessment modality (Table 2), based on a treatment hypothesised to reduce the rate of change in each outcome by 50% (90% power and 5% significance level). Based on these results, recommendations for selecting biomarkers for short PoC studies and longer-term Phase III trials are provided as "ticks" (show potential), "crosses" (unlikely to be suitable) and "question marks" (further data is required due to wide confidence intervals). An important caveat of this figure is that sample sizes are heavily dependent on the magnitude of the hypothesised treatment effect (Figure 1). For example, requirements would be four times larger if the effect was reduced to 25%. Nevertheless, this approach does provide an estimate of sample-size requirements to sufficiently power trials, as well as a means of comparing the outcomes across assessment modalities.

For example, in order to detect therapeutic effects on ventricular expansion following treatment periods of 6-, 9- or 15-months, sample-size requirements per treatment arm would be 134 (95% CI: 64, 495), 98 (95% CI: 51, 275) and 80 (95% CI: 48, 186) respectively, for 50% efficacy. Considering the magnitude of the sample sizes and the width of the confidence intervals, ventricular expansion may be a suitable biomarker for use in short-term PoC studies, as well as trials over a longer duration (Figure 2).

Conversely, to assess the effect of a therapy on motor progression, the commonly-applied UHDRS-TMS may be suitable for use over 9- and 15-month intervals, given a

50% treatment effect; however, the wide confidence intervals around these sample sizes indicate a lack of precision (Figure 2).

Generalizability

It is important to note that observational data should only be used to inform clinical trials involving similar cohorts and observational periods. The current study focussed predominately on stage 1 HD, the very early clinical phase of the disease, since disease-modifying treatments are most likely to be efficacious in preserving function and quality of life when administered at this point. Therapies shown to be effective in these cohorts within an acceptable safety profile, may be administered during the premanifest stages of the disease, prior to clinical onset. The observational PREDICT-HD study, which focuses on the premanifest stages of the disease, is ideally positioned to inform the design of such trials.[8]

Limitations

We must acknowledge the potential limitations of using neuroimaging biomarkers as efficacy readouts. It is possible that a positive macrostructural neuroimaging readout over six or nine months may not be indicative of longer-term clinical or functional improvement. Although associations between change in neuroimaging measures and functional decline have been reported in HD, causality is yet to be demonstrated.[4,7] Furthermore, these readouts may not be suitable for all types of intervention; their utility may be dependent on the mechanism-of-action of the therapy, together with the time required for it to mediate an effect. Nevertheless, these neuroimaging measures are able to track the progression of pathological atrophy over short time intervals,

reproducible across multiple sites and objective. They may provide valuable biomarkers in the assessment of disease-modifying compounds. Another limitation includes the decision to focus on the corpus callosum as a whole, when there is evidence that each sub-region of the corpus callosum projects to distinct cortical regions and is likely to be differentially implicated in the disease process. Future work should investigate these sub-structures independently, and whether the added complexity of delineating smaller with less well-defined regions is offset by a stronger atrophy signal.

None of the participants in the current study were enrolled in clinical trials; however, many were on medications which target the central nervous system (CNS) (Supplemental Table 2). Mean dosages of CNS-targeting drugs were relatively low, with overlap in usage between groups. This study was not designed to examine the specific effects of medication on each outcome; however, we acknowledge medication usage as a potential confounder.

Conclusion

The short-interval observational data presented here are complimentary to findings over longer intervals in others such as the TRACK-HD and the PREDICT-HD studies. Taken together, these studies can provide data to directly inform the design of clinical trials in HD, facilitating the evaluation of treatments designed to slow the course of this devastating disease. Since HD is often regarded as a model neurodegenerative disease, amenable to early intervention,[1] research into this disorder may inform early-intervention strategies for more prevalent neurodegenerative diseases.

ACKNOWLEDGEMENT

The authors would like to thank the patients and controls who took part in this study, along with all the PADDINGTON study Work Package 2 site staff at Paris, Leiden, Ulm and London, and Eileanoir Johnson for her assistance with the medication table. This work has been supported by the European Union – PADDINGTON project, contract no. HEALTH-F2-2010-261358. RIS is supported by the CHDI/High Q Foundation, a not for profit organization dedicated to finding treatments for Huntington's disease. This work was undertaken at UCLH/UCL which received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. SJT acknowledges support of the National Institute for Health Research through the Dementias and Neurodegenerative Research Network, DeNDRoN.

PADDINGTON Work-Package 2 contributors:

Netherlands: Dr Ellen 't Hart MSc, Verena Rödig MSc, Anne Schoonderbeek MSc (Leiden University of Medical Sciences). **UK:** Victoria Perry BSc, Nicola Robertson BSc (UCL Institute of Neurology, London). **France:** Dr Perrine Charles MD PhD, Dr Claire Ewenczyk MD, Dr Stephan Klebe MD (Assistance Publique-Hôpitaux de Paris, Paris), Dr Damien Justo PhD (Université Pierre et Marie Curie, Paris). **Germany:** Sabrina Betz, Dr Jens Dreyhaupt PhD, Carolin Eschenbach, Ms Jeton Iseni, Daniela Schwenk, Dr Michael Orth Associate Professor MD, Sonja Trautmann, Nurse, Ms Karin Schiefele, Irina Blankin BSc, Ms Theresia Kelm BSc, Rosine Scherer Diploma, Felix Mudoh Tita MSc, Katja Vitkin (Ulm University). **Italy:** Dr Giovanna Tripepi PhD, Dr Giuseppe Pollio PhD.

The authors have no conflict of interests to declare.

FIGURE LEGENDS

Figure 1 Relationship between effect sizes and sample-size requirements for randomised controlled trials where the outcome is a change measure between two time points. Plots of this relationship are shown for treatments with efficacy levels of 25% (red), 50% (blue) and 100% (green), assuming 90% power and a 5% significance level.

Figure 2 Suggested biomarker selection for trials of a 50% effective disease-modifying agent Sample-size requirements are per treatment arm; calculated using the standard formula[22], with 90% power and two-tailed $p<0.05$, for therapies with 50% estimated treatment efficacy. Recommendations are given as ticks ("show potential"), crosses ("unlikely to be suitable") and question marks (“further data required – wide confidence intervals”).

REFERENCES

1 Ross CA and Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurology 2011;10:83-98.

2 Berger JR, Choi D, Kaminski HJ *et al.* Importance and Hurdles to Drug Discovery for Neurological Disease. Annals of Neurology 2013;74:441-446.

3 Dominguez JF, Egan GF, Gray MA *et al.* Multi-Modal Neuroimaging in Premanifest and Early Huntington's Disease: 18 Month Longitudinal Data from the IMAGE-HD Study. Plos One 2013;8.

4 Tabrizi SJ, Reilmann R, Roos RA *et al.* Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. Lancet Neurology 2012;11:42-53.

5 Tabrizi SJ, Scahill RI, Durr A *et al.* Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. Lancet Neurology 2012;10:31-42.

6 Tabrizi SJ, Langbehn DR, Leavitt BR *et al.* Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. Lancet Neurology 2009;8:791-801.

7 Tabrizi SJ, Scahill RI, Owen G *et al.* Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. Lancet Neurology 2013;12:637-649.

8 Paulsen JS, Hayden M, Stout JC *et al.* Preparing for preventive clinical trials: the Predict-HD study. Arch. Neurol. 2006;63:883-890.

9 Aylward EH, Nopoulos PC, Ross CA *et al.* Longitudinal change in regional brain volumes in prodromal Huntington disease. J. Neurol. Neurosurg. Psychiatry 2011;82:405-410.

10 Hobbs NZ, Cole JH, Farmer RE *et al.* Evaluation of multi-modal, multi-site neuroimaging measures in Huntington's disease: Baseline results from the PADDINGTON study. Neuroimage: Clinical 2013;2:204-211.

11 Huntington’s Disease Study Group. Unified Huntington's disease rating scale: Reliability and consistency. Movement Disorders 1996;11:136-142.

12 Muller HP, Kassubek J, Gron G *et al.* Evaluating multicenter DTI data in Huntington's disease on site specific effects: An ex post facto approach. Neuroimage:Clinical 2013;2:161-167.

13 Freeborough PA, Fox NC and Kitney RI. Interactive algorithms for the segmentation and quantitation of 3-D MRI brain scans. Computer Methods and Programs in Biomedicine 1997;53:15-25.

14 Freeborough PA and Fox NC. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *I*EEE Trans. Med. Imaging 1997;16:623-629.

15 Leung KK, Clarkson MJ, Bartlett JW *et al.* Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. Neuroimage. 2010;50:516-523.

16 Magnotta VA, Harris G, Andreasen NC, O'Leary DS, Yuh WT and Heckel D. (2002) Structural MR image processing using the BRAINS2 toolbox. Comput. Med. Imaging Graph. 2002;26:251-264.

17 Hobbs NZ, Henley SM, Ridgway GR. *et al.* The progression of regional atrophy in premanifest and early Huntington's disease: a longitudinal voxel-based morphometry study. J Neurol. Neurosurg. Psychiatry 2010;81:756-763.

18 Christensen GE, Rabbitt RD and Miller MI. Deformable templates using large deformation kinematics. IEEE Transactions on Image Processing 1996; 5:1435-1447.

19 Reuter M, Schmansky NJ, Rosas HD and Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 2012;61:1402-1418.

20 Desikan RS, Segonne F, Fischl B *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31:968-980.

21 Rubin DB. Interference and missing data. Biometrika 1976; 63:581-592.

22 Julious SA. Sample sizes in clinical trials, Boca Raton: Chapman and Hall; 2009.

23 Carpenter J and Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 2000;19:1141-1164.

24 Rothman KJ. No adjustments are needed for Multiple Comparisons. Epidemiology 1990;1:43-46.

25 Penney JB, Vonsattel JP, MacDonald ME, Gusella JF and Myers RH. CAG repeat number governs the development rate of pathology in Huntington's disease. Annals of Neurology 1997;41:689-692.

26 Rosas HD, Doros G, Gevorkian S *et al.* PRECREST: A phase II prevention and biomarker trial of creatine in at-risk Huntington's Disease. Neurology 2014;82:1-8.

27 Stout JC, Jones R, Labuschagne I *et al.* Evaluation of longitudinal 12 and 24 month cognitive outcomes in premanifest and early Huntington's disease. J Neurol Neurosurg Psychiatry 2012;83:687-694.