

RESEARCH PAPER

Evaluation of longitudinal 12 and 24 month cognitive outcomes in premanifest and early Huntington's disease

Julie C Stout,¹ Rebecca Jones,² Izelle Labuschagne,¹ Alison M O'Regan,¹ Miranda J Say,³ Eve M Dumas,⁴ Sarah Queller,^{1,5} Damian Justo,⁶ Rachelle Dar Santos,⁷ Allison Coleman,⁷ Ellen P Hart,⁴ Alexandra Dürr,⁶ Blair R Leavitt,⁷ Raymund A Roos,⁴ Doug R Langbehn,⁸ Sarah J Tabrizi,³ Chris Frost²

► Additional materials are published online only. To view these files please visit the journal online (<http://jnnp.bmj.com/content/83/7.toc>).

¹School of Psychology and Psychiatry, Monash University, Melbourne, Victoria, Australia

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

³UCL Institute of Neurology, University College London, London, UK

⁴Department of Neurology, Leiden University Medical Centre, The Netherlands

⁵Queller Consulting, Dunedin, Florida, USA

⁶Department of Genetics and Cytogenetics, and INSERM UMR S679, APHP Hôpital de la Salpêtrière, Paris, France

⁷Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

⁸Departments of Psychiatry and Biostatistics (Secondary), University of Iowa, Iowa City, Iowa, USA

Correspondence to

Professor J C Stout, School of Psychology and Psychiatry, Monash University, Wellington Road, Melbourne, Victoria 3800, Australia; julie.stout@monash.edu

Received 1 December 2011

Revised 22 March 2012

Accepted 27 March 2012

Published Online First

7 May 2012



This paper is freely available online under the BMJ Journals unlocked scheme, see <http://jnnp.bmj.com/site/about/unlocked.xhtml>

ABSTRACT

Background Deterioration of cognitive functioning is a debilitating symptom in many neurodegenerative diseases, such as Huntington's disease (HD). To date, there are no effective treatments for the cognitive problems associated with HD. Cognitive assessment outcomes will have a central role in the efforts to develop treatments to delay onset or slow the progression of the disease. The TRACK-HD study was designed to build a rational basis for the selection of cognitive outcomes for HD clinical trials.

Methods There were a total of 349 participants, including controls (n=116), premanifest HD (n=117) and early HD (n=116). A standardised cognitive assessment battery (including nine cognitive tests comprising 12 outcome measures) was administered at baseline, and at 12 and 24 months, and consisted of a combination of paper and pencil and computerised tasks selected to be sensitive to cortical-striatal damage or HD. Each cognitive outcome was analysed separately using a generalised least squares regression model. Results are expressed as effect sizes to permit comparisons between tasks.

Results 10 of the 12 cognitive outcomes showed evidence of deterioration in the early HD group, relative to controls, over 24 months, with greatest sensitivity in Symbol Digit, Circle Tracing direct and indirect, and Stroop word reading. In contrast, there was very little evidence of deterioration in the premanifest HD group relative to controls.

Conclusions The findings describe tests that are sensitive to longitudinal cognitive change in HD and elucidate important considerations for selecting cognitive outcomes for clinical trials of compounds aimed at ameliorating cognitive decline in HD.

INTRODUCTION

Cognitive decline is a serious debilitating symptom in neurodegenerative diseases, resulting in untold suffering and huge financial costs. Thus treatments for cognitive decline are urgently needed. These potential treatments fall into two broad categories: (a) disease modifying treatments, which are aimed at changing the neuropathological progression (eg, halting, slowing); and (b) symptom focused treatments, which are aimed at enhancing the function of compromised neural systems. Although

symptom focused treatments, such as the use of cholinesterase inhibitors in Alzheimer's and other diseases, have met with a moderate degree of success, there are, as of yet, no disease modifying treatments for any neurodegenerative disease.

Huntington's disease (HD) is a fully penetrant, autosomal dominant neurodegenerative disease. Unlike Alzheimer's disease or Parkinson's disease, for which the genetic risk factors are far less predictive, it is possible to know with certainty who will develop HD far in advance of the symptoms and signs of disease. As such, HD has emerged from the neurodegenerative diseases as a potential opportunity for the development of the first disease modifying intervention strategies. People who have the HD CAG expansion usually start life functioning normally and then begin to gradually develop involuntary movements, psychiatric symptoms and cognitive decline, eventually leading to death typically 15–20 years following diagnosis.^{1,2} As potential interventional compounds are identified, it will theoretically be possible to identify people at risk who can be treated preventatively, in the premanifest period, to impede the development of disease signs and symptoms. However, because it is essential to be able to test intervention strategies in trials of reasonably limited duration, the slowness with which HD progresses in the premanifest period is prohibitive, and instead it will be necessary to test for drug effects in already diagnosed patients when progression may be rapid enough to get efficient readouts from clinical trials.

For any disease with progressive cognitive decline, success in finding treatments to prevent or slow cognitive deterioration rests on the availability of cognitive outcomes that are tolerable in the clinical trial setting and are responsive to treatment. Generally, cost considerations mean that clinical trial duration is limited to 1 or 2 years at most, and sample sizes must be in the low hundreds rather than in the thousands. Thus clinical trials for cognitive interventions are fully reliant on the availability of cognitive outcome measures that can reveal change within this interval. At this time, there is no currently accepted battery of cognitive tests—that is, ready for clinical trials—in either diagnosed or premanifest HD.

This is the first report to examine longitudinal 12 and 24 month progression in late premanifest and early HD compared with controls, with respect to feasibility and methodology for these cognitive measures for HD trials, although a subset of the 12 and 24 month data was previously reported.^{3,4} The cognitive battery was administered in the context of TRACK-HD, a multisite, observational, longitudinal study aimed at identifying biological and clinical markers in premanifest and early HD individuals, across domains of cognition, psychiatry, quantitative motor and neuroimaging. The aims of the analyses reported here were to: (a) determine whether progression of cognitive decline could be detected at 12 or 24 month intervals (to approximate feasible timelines of future clinical trials); (b) quantify the effect sizes (ES) for rates of change in cognition in order to facilitate power calculations for future trials; and (c) determine whether particular cognitive measures show statistically significant superiority over other measures in the ability to detect change over time. For future clinical trials, cognitive measures that require the smallest sample sizes for any chosen treatment effect will be those with the largest ES. For this reason, and also to better understand how the ubiquitous practice effects present in cognitive assessment are exhibited in premanifest diagnosed groups compared with disease-free participants, we included a disease-free comparison group in our analyses. For the purposes of sample size calculations, we consider a 100% effective treatment to be one where the mean change in a treated group is the same as that in the disease-free group.

METHODS

Participants

Briefly, participants were recruited from four sites, including Vancouver, Paris, Leiden and London, as part of the TRACK-HD study.⁵ Participants were 18–65 years old, able to tolerate and safely undergo magnetic resonance imaging (MRI), were not participants in a clinical drug trial and were free of concomitant other major neurological, psychiatric or medical illnesses (including significant head injury, drug/alcohol abuse). Inclusion in the premanifest group was defined at study entry by a disease burden score of $>250^6$ and a total motor score ≤ 5 , as assessed by the motor examination of the United Huntington's Disease Rating Scale (UHDRS-99).⁷ The early HD group included individuals at stages 1 or 2 according to the UHDRS Total Functional Capacity score at the baseline assessment. Controls were primarily spouses or partners and gene negative siblings to maximise consistency of environments. Where possible, groups were frequency matched (ie, having similar distributions) on age, sex and education; as expected, given the progressive nature of HD, the early HD group was slightly older than the premanifest and control groups (see table 1). A total of 366 participants were enrolled at baseline. Here we report on a total of 349 participants (see supplement for more detail, available online only).

Cognitive assessment

Table 2 provides a list of the cognitive tasks and the variables analysed for this study, with details of the cognitive methods for each test presented in the supplement (available online only). Briefly, examiners were trained in person by the first author for standardised test administration of a set of paper and pencil and computerised tasks, and then they tested participants in the language spoken locally at each site (French, Dutch and English) as part of an annual TRACK-HD visit. Here we report on nine tests (12 primary outcomes) that were administered at all three visits (0, 12 and 24 months).

Table 1 Summary of participant characteristics

	Control	Premanifest HD*	Early HD
No of participants†	116	117	116
Age (years)‡			
Mean (SD)	46.2 (10.2)	40.8 (8.9)	49.2 (9.7)
Range	23.0–65.7	18.6–64.1	22.8–64.1
Gender			
Women (n (%))	65 (56.0)	64 (54.7)	63 (54.3)
Men (n (%))	51 (44.0)	53 (45.3)	53 (45.7)
Education§			
Mean (SD)	4.0 (1.3)	4.0 (1.2)	3.7 (1.3)
CAG length¶			
Mean (SD)	—	43.1 (2.4)	43.6 (3.0)
Range	—	39–52	39–59
Disease burden score‡ **			
Mean (SD)	—	293.8 (47.6)	375.5 (74.3)

*This group had an estimated median of 10.8 years to onset.

†Number of participants in the TRACK-HD study with at least one follow-up measure on at least one of the 12 cognitive tasks featured in this paper.

‡Age and disease burden score as measured at baseline.

§Education level was reported according to the ISCED education classification system.⁸

¶Two premanifest and three early HD participants had CAG repeats of 39, the remainder were all ≥ 39 .

**Disease burden score = age \times (CAG length–35.5).

HD, Huntington's disease.

Statistical methods

Cognitive outcomes were analysed separately using a generalised least squares regression model for repeated measures of the outcome at baseline, and at 12 and 24 months (additional details in the supplement, available online only). For a given outcome, participants were excluded from data analysis if they had data at only one of the three visits. ES for differences in the rate of change observed over both 12 and 24 months for each task were calculated as the estimated difference in longitudinal change in each disease group relative to controls, divided by the residual SD of change in the disease group. To compare ES magnitudes between the 12 cognitive outcomes, we calculated differences between ES for each pair of tasks for both the 12 and 24 month change. We estimated 95% CIs for the ES and pairwise ES differences using the bias corrected and accelerated bootstrap method with 2000 replications.⁹ All analyses were performed using SAS V.9.2. (Stata Corporation).

RESULTS

All 12 of the cognitive outcomes showed evidence of deterioration in the early HD group, relative to controls, over 24 months. Differences were statistically significant ($p < 0.05$) for all measures except Trails B and 1.8 Hz Paced Tapping, which were borderline statistically significant ($0.05 < p < 0.1$). In contrast, very little evidence of decline was detectable in the premanifest group. Table 3 presents the unadjusted means at baseline, and at 12 and 24 months, and table 4 displays the adjusted means between group differences in longitudinal change.

Despite the consistent evidence of deterioration in the early HD group, the way this deterioration was expressed varied. For example, in some tasks, the early HD group showed a *decline* in cognitive performance at subsequent visits whereas the control group showed *improvements* (ie, practice effects), resulting in large longitudinal differences between groups in rates of change. The Symbol Digit Modalities Test (SDMT) showed this pattern; in early HD there was a decline from 33.9 at baseline to 31.0 at 24 months compared with controls who improved from 52.4 at baseline to 54.4 at 24 months. After adjustment for demographic

Table 2 Cognitive test battery information

Task	Primary variable	Cognitive domain	Task administration times* (min:s)					
			Controls		Premanifest HD		Early HD	
			Mean	Longest†	Mean	Longest†	Mean	Longest†
SDMT‡	Number correct	Psychomotor speed, working memory	2:51	5:00	3:21	5:00	3:14	5:00
Stroop Word Reading‡	Number correct	Psychomotor, speeded word reading	2:03	4:00	2:18	4:00	2:02	5:00
Trails A‡	Completion time (s)	Attention, psychomotor processing	1:42	4:00	1:54	3:00	2:09	4:00
Trails B‡	Completion time (s)	Attention, set shifting, psychomotor processing	2:23	8:00	2:43	5:00	3:41	12:00
Paced Tapping (1.8 and 3 Hz)§	Precision (1/SD of ITI in 1/ms)	Psychomotor, movement timing (slow and fast)	6:53	10:00	7:24	10:00	7:17	12:00
Serials 2 s with tapping§	Number correct subtractions	Psychomotor, speed, dexterity with cognitive load	5:20	10:00	5:35	9:00	5:47	8:00
Spot the Change set size 5§	Number correct adjusted for guessing (k)	Visual working memory	7:32	11:00	8:17	11:00	7:35	10:00
Emotion Recognition§	Number correct combined negative emotions	Perceptual (sensory) emotion recognition	7:04	12:00	7:58	12:00	9:24	14:00
UPSIT¶ ‡	Number correct	Odour recognition	NA	NA	NA	NA	NA	NA
Circle Tracing direct and indirect§	Annulus length (log cm)	Motor speed, planning and correction	8:51	15:00	9:22	13:00	9:30	13:00
		Total time**	44:65	79:00	48:86	72:00	50:65	83:00

*Times for each outcome on the tasks are based on the commencement of the standard operating procedures (SOP) until the end of the actual task time.

†The maximum time required by any participant to complete the task (SOP time plus actual task time).

‡Pencil and paper tasks.

§Computerised tasks.

¶Time of administration of the UPSIT (the University of Pennsylvania Smell Identification Test) was not available (NA).

**The total time of the battery after adding up all the longest times (across various participants) on the individual tasks.

HD, Huntington's disease; ITI, inter-trial interval; SDMT, Symbol Digit Modalities Test.

factors, the early HD group relative to controls declined by 2.63 points (95% CI 1.91 to 3.34) per year over 24 months. A similar pattern was observed on the Stroop Test Word Reading condition, with the early HD group declining 4.21 points (95% CI 2.78 to 5.65) per year more than controls over 24 months. On other tests, such as the Circle Tracing indirect condition, both controls and the early HD group exhibited practice effects but this effect was markedly greater in controls, indicating a *relative deterioration* in the early HD group. For example, for Circle Tracing indirect, controls improved from 5.59 at baseline to 6.16 at 24 months whereas the early HD group improved only from 5.14 at baseline to 5.38 at 24 months. Circle Tracing direct showed a similar pattern. Finally, in some tasks, such as the the University of Pennsylvania Smell Identification Test (UPSIT), the early HD group *declined* while the control group's performance stayed *stable*; controls scored 17.16 at baseline and 17.13 on average at 24 months whereas early HD scored 13.51 at baseline and 12.59 at 24 months. After adjustment, performance of the early HD group compared with controls decreased by 0.52 points (95% CI 0.21 to 0.83; $p=0.001$) per year over 24 months.

The strongest and most consistent evidence of differences in longitudinal rates of change in the early HD group compared with controls, as indicated by large standardised ES, were in three outcomes which showed significant effects at 12 and 24 months (all p 's < 0.0005). As an illustration of these, 24 month ES for differences from controls were SDMT=1.00 (95% CI 0.70 to 1.30), Circle Tracing Indirect=0.85 (95% CI 0.58 to 1.18) and Stroop Word Reading=0.73 (95% CI 0.48 to 1.03). In contrast, for other cognitive outcome measures, such as Negative Emotion Recognition and Spot the Change (visual working memory), we observed strong evidence only at 24 months (emotion ES: 0.49; 95% CI 0.21 to 0.77; $p=0.0003$; spot ES: 0.40; 95% CI 0.16 to 0.68; $p=0.0025$) whereas at 12 months evidence of faster rates of decline in early HD was only weak (emotion ES: 0.27; 95% CI 0.03 to 0.52; $p=0.034$; spot ES: 0.23; 95% CI -0.03 to 0.46; $p=0.070$). Finally, some tasks, including 1.8 Hz

Paced Tapping and Trails B, showed no statistically significant deterioration over 12 months ($p>0.50$) and only weak evidence of decline over 24 months (1.8 Hz Tapping ES: 0.32; 95% CI -0.02 to 0.74; $p=0.070$; Trails B ES: 0.19; 95% CI 0.00 to 0.39; $p=0.067$). See table 5 for full details of ES for all outcomes.

In contrast with the clear evidence of decline in early HD, we found very little evidence of measurable deterioration in the premanifest group relative to controls over either 12 or 24 months. The strongest suggestions of longitudinal decline in the premanifest group came from the Circle Tracing indirect condition and SDMT, with ES of 0.23 (95% CI -0.05 to 0.51) and 0.20 (95% CI -0.03 to 0.43), respectively, over 12 months and 0.19 (95% CI -0.10 to 0.48) and 0.14 (95% CI -0.11 to 0.38) over 24 months. None of these longitudinal effects reached the statistical significance threshold of $p<0.05$.

To facilitate more robust comparisons between tasks, we examined whether some tasks were statistically superior to other tasks in detecting longitudinal changes. Results of these analyses indicated that, whereas in absolute terms the SDMT had larger ES at both 12 and 24 months compared with other cognitive tasks, the SDMT ES were not statistically significantly larger than many other tests. More specifically, SDMT was not significantly better at detecting longitudinal differences between early HD patients and controls than the Circle Tracing indirect condition, Stroop Word Reading or 3 Hz Paced Tapping, for either the 12 or 24 month time periods. Neither was there any evidence that Trails B, the task with the smallest absolute ES, was significantly worse than Negative Emotion Recognition, Spot the Change, Trails A or Paced Tapping at either 1.8 or 3 Hz. We were thus unable to distinguish either a single 'best' or a 'worst' performing test within the cognitive battery on the basis of ES differences. See table 6 for full results of comparisons of ES between outcomes.

An important caveat for reconciling the results presented here with our previous reports on Circle Tracing tasks at the 12 month time point is that in the current analyses we have taken

Table 3 Summary of performance on cognitive assessments at baseline, and after 12 months and 24 months of follow-up

Assessment	Controls			Premanifest HD			Early HD		
	Baseline	12 months	24 months	Baseline	12 months	24 months	Baseline	12 months	24 months
SDMT (number correct)*									
Mean (SD)	52.41 (9.41)	54.50 (9.00)	54.37 (9.86)	51.46 (10.29)	52.51 (10.28)	52.03 (10.88)	33.86 (10.11)	32.25 (10.79)	31.01 (11.23)
N	116	116	110	116	116	111	116	113	109
Stroop Word Reading (number correct)*									
Mean (SD)	106.09 (16.47)	107.85 (15.99)	107.27 (16.03)	99.76 (16.16)	101.57 (16.94)	99.46 (15.58)	78.36 (19.56)	75.56 (20.37)	71.97 (22.11)
N	116	116	110	116	116	111	116	113	111
Trails A completion time (s)†									
Mean (SD)	27.72 (10.29)	24.92 (8.46)	26.16 (10.76)	28.58 (9.90)	27.47 (10.56)	26.96 (8.79)	50.19 (23.52)	50.67 (24.39)	54.41 (32.13)
N	116	116	110	116	116	111	116	113	111
Trails B completion time (s)†									
Mean (SD)	61.03 (25.74)	60.32 (32.58)	57.18 (24.35)	66.86 (28.48)	63.83 (28.69)	66.45 (31.61)	139.62 (74.94)	139.15 (70.70)	143.2 (78.97)
N	116	116	110	116	116	111	115	113	110
Paced Tapping 3 Hz precision (1/SD of ITI in 1/ms)†									
Mean (SD)	0.025 (0.011)	0.027 (0.010)	0.027 (0.011)	0.020 (0.008)	0.022 (0.009)	0.020 (0.008)	0.010 (0.007)	0.010 (0.007)	0.009 (0.006)
N	116	114	110	116	116	111	112	110	104
Paced Tapping 1.8 Hz precision (1/SD of ITI in 1/ms)†									
Mean (SD)	0.022 (0.008)	0.023 (0.008)	0.023 (0.007)	0.018 (0.007)	0.019 (0.007)	0.018 (0.006)	0.009 (0.006)	0.010 (0.005)	0.009 (0.006)
N	116	115	110	116	117	111	112	110	104
Serial 2 s with tapping (correct subtractions)*									
Mean (SD)	9.44 (2.45)	9.50 (2.57)	10.22 (2.93)	8.22 (2.71)	8.12 (2.80)	8.69 (2.90)	5.94 (2.29)	5.58 (2.58)	6.16 (2.64)
N	115	116	109	117	117	111	107	105	98
Spot the Change set size 5 (k)*									
Mean (SD)	3.09 (1.11)	3.09 (1.23)	3.34 (1.12)	2.84 (1.32)	2.91 (1.18)	2.87 (1.09)	1.65 (1.27)	1.38 (1.51)	1.44 (1.44)
N	114	115	110	116	115	111	107	106	104
Negative Emotion Recognition (number correct)*									
Mean (SD)	26.02 (5.01)	26.90 (5.21)	27.30 (5.88)	23.48 (5.71)	24.70 (5.66)	24.67 (6.26)	16.42 (6.46)	16.33 (6.80)	15.71 (7.22)
N	110	115	110	111	116	111	105	110	105
UPSIT (number correct out of 20)*									
Mean (SD)	17.16 (2.21)	16.99 (1.98)	17.13 (2.29)	16.58 (2.56)	16.70 (2.21)	16.50 (2.26)	13.51 (3.26)	12.65 (3.28)	12.59 (3.66)
N	116	116	107	117	116	111	116	113	108
Circle Tracing direct annulus length (log cm)*									
Mean (SD)	6.84 (0.36)	7.05 (0.31)	7.18 (0.28)	6.73 (0.37)	6.94 (0.35)	7.05 (0.31)	6.38 (0.41)	6.54 (0.41)	6.50 (0.45)
N	115	115	109	116	116	111	112	112	106
Circle Tracing indirect annulus length (log cm)*									
Mean (SD)	5.59 (0.41)	6.03 (0.40)	6.16 (0.43)	5.56 (0.43)	5.90 (0.39)	6.03 (0.38)	5.14 (0.48)	5.38 (0.52)	5.38 (0.49)
N	115	116	109	116	116	111	112	112	106

*Higher values mean better performance.

†Lower values mean better performance.

HD, Huntington's disease; ITI, inter-trial interval; SDMT, Symbol Digit Modalities Test; UPSIT, the University of Pennsylvania Smell Identification Test.

Table 4 Between group differences in annualised rate of longitudinal change adjusted for age, sex, centre and education

Assessment	Premanifest HD versus controls		Early HD versus controls	
	12 months	24 months	12 months	24 months
SDMT (number correct)				
Adjusted difference (95% CI)	-1.17 (-2.56 to 0.23)	-0.49 (-1.32 to 0.34)	-3.62 (-4.86 to -2.38)	-2.63 (-3.34 to -1.91)
p Value	0.10	0.25	<0.0001	<0.0001
Stroop Word Reading (number correct)				
Adjusted difference (95% CI)	0.14 (-2.43 to 2.72)	-0.74 (-2.07 to 0.60)	-4.84 (-7.33 to -2.35)	-4.21 (-5.65 to -2.78)
p Value	0.91	0.28	0.0002	<0.0001
Trails A completion time (s)				
Adjusted difference (95% CI)	1.44 (-0.63 to 3.50)	-0.10 (-1.27 to 1.07)	3.25 (-0.36 to 6.85)	3.16 (0.86 to 5.45)
p Value	0.17	0.87	0.077	0.0073
Trails B completion time (s)				
Adjusted difference (95% CI)	-1.67 (-6.73 to 3.39)	1.55 (-1.09 to 4.19)	-1.95 (-10.67 to 6.76)	4.66 (-0.34 to 9.65)
p Value	0.52	0.25	0.66	0.067
Paced Tapping 3 Hz precision (1/SD of ITI in 1/ms)				
Adjusted difference (95% CI)	-0.0009 (-0.0033 to 0.0016)	-0.0007 (-0.0020 to 0.0006)	-0.0022 (-0.0044 to 0.0000)	-0.0013 (-0.0025 to -0.0001)
p Value	0.49	0.27	0.046	0.032
Paced Tapping 1.8 Hz precision (1/SD of ITI in 1/ms)				
Adjusted difference (95% CI)	0.0004 (-0.0012 to 0.0021)	-0.0003 (-0.0012 to 0.0005)	0.0002 (-0.0013 to 0.0017)	-0.0007 (-0.0015 to 0.0001)
p Value	0.60	0.46	0.79	0.070
Serial 2 s with tapping (correct subtractions)				
Adjusted difference (95% CI)	-0.09 (-0.40 to 0.21)	-0.11 (-0.31 to 0.10)	-0.39 (-0.69 to -0.09)	-0.38 (-0.59 to -0.18)
p Value	0.55	0.30	0.012	0.0003
Spot the Change set size 5 (k)				
Adjusted difference (95% CI)	0.07 (-0.24 to 0.38)	-0.10 (-0.26 to 0.05)	-0.29 (-0.60 to 0.02)	-0.25 (-0.41 to -0.09)
p Value	0.66	0.20	0.070	0.0025
Negative Emotion Recognition (number correct)				
Adjusted difference (95% CI)	0.09 (-0.90 to 1.09)	-0.17 (-0.75 to 0.42)	-1.15 (-2.22 to -0.09)	-1.13 (-1.74 to -0.53)
p Value	0.85	0.57	0.034	0.0003
UPSIT (number correct out of 20)				
Adjusted difference (95% CI)	0.30 (-0.15 to 0.76)	0.01 (-0.23 to 0.25)	-0.76 (-1.32 to -0.20)	-0.52 (-0.83 to -0.21)
p Value	0.19	0.95	0.0086	0.0010
Circle Tracing direct annulus length (log cm)				
Adjusted difference (95% CI)	0.003 (-0.088 to 0.093)	-0.017 (-0.057 to 0.024)	-0.057 (-0.151 to 0.037)	-0.103 (-0.146 to -0.061)
p Value	0.96	0.42	0.23	<0.0001
Circle Tracing indirect annulus length (log cm)				
Adjusted difference (95% CI)	-0.081 (-0.176 to 0.013)	-0.033 (-0.083 to 0.016)	-0.217 (-0.319 to -0.115)	-0.180 (-0.234 to -0.127)
p Value	0.091	0.18	<0.0001	<0.0001

HD, Huntington's disease; ITI, inter-trial interval; SDMT, Symbol Digit Modalities Test; UPSIT, the University of Pennsylvania Smell Identification Test.

care to avoid inflation of longitudinal ES that arise due to a combination of large baseline differences between groups and an association between change and baseline performance. Specifically, because changes tend to be smaller in cases with lower baseline levels (ie, HD cases), it is implausible that even a 100% effective treatment will render the mean change in outcome in the HD group to be as great as that in the control group, and hence the ES will be unrealistically large for the purposes of estimating sample sizes for clinical trials. For this reason, we logarithmically transformed the Circle Tracing data as this removed any dependency of change on baseline, as assessed by testing for associations between change and mean levels.¹⁰

DISCUSSION

In this study, we found highly consistent evidence that longitudinal cognitive decline is detectable across a 24 month interval in early HD. Changes in 10 of the 12 cognitive outcome measures, which were derived from nine distinct cognitive tests, were statistically significant compared with controls, with medium to large ES. About half of the cognitive measures also

showed statistically significant (small to medium) effects after only 12 months of follow-up. In contrast with the early HD findings, we did not detect statistically significant longitudinal decline at either 12 or 24 months in the premanifest sample relative to change in controls. Because we studied sample sizes and an overall duration of follow-up relevant to clinical trials, as well as including both premanifest and early HD participants in the study, the ES results from this study are useful for clinical trial planning in HD. Thus these results provide ample cognitive outcomes sensitive at 12 or 24 months in early HD, indicating that it is now possible to conduct treatment trials aimed at slowing cognitive deterioration in early stage patients.

In contrast with the findings in early HD, our results indicate that for premanifest HD, rates of progression of these cognitive outcomes appear to be too slow to detect with a reasonable sample size in a time period reasonable for a clinical trial. Importantly, the lack of significant findings in premanifest HD does not mean an absence of progressive cognitive decline throughout the premanifest HD period. Rather, it seems more likely that the magnitude of this decline is too small and/or the rate of progression is too slow to be detected over 24 months in

Table 5 Standardised effect sizes of between group differences in change adjusted for age, sex, centre and education

Assessment	Premanifest HD versus controls		Early HD versus controls	
	12 months	24 months	12 months	24 months
SDMT (number correct)				
Effect size (95% CI)	0.20 (−0.03 to 0.43)	0.14 (−0.11 to 0.38)	0.75 (0.51 to 1.06)	1.00 (0.70 to 1.30)
Stroop Word Reading (number correct)				
Effect size (95% CI)	−0.01 (−0.22 to 0.24)	0.15 (−0.11 to 0.43)	0.50 (0.23 to 0.79)	0.73 (0.48 to 1.03)
Trails A completion time (s)				
Effect size (95% CI)	0.19 (−0.07 to 0.48)	−0.02 (−0.29 to 0.27)	0.18 (−0.03 to 0.41)	0.28 (0.10 to 0.44)
Trails B completion time (s)				
Effect size (95% CI)	−0.09 (−0.39 to 0.18)	0.15 (−0.12 to 0.39)	−0.05 (−0.25 to 0.18)	0.19 (0.00 to 0.39)
Paced Tapping 3 Hz precision (1/SD of ITI in 1/ms)				
Effect size (95% CI)	0.11 (−0.20 to 0.42)	0.19 (−0.14 to 0.53)	0.48 (0.01 to 1.06)	0.49 (0.07 to 1.01)
Paced Tapping 1.8 Hz precision (1/SD of ITI in 1/ms)				
Effect size (95% CI)	−0.07 (−0.32 to 0.17)	0.11 (−0.20 to 0.40)	−0.04 (−0.34 to 0.25)	0.32 (−0.02 to 0.74)
Serial 2 s with tapping (correct subtractions)				
Effect size (95% CI)	0.07 (−0.15 to 0.32)	0.14 (−0.13 to 0.42)	0.33 (0.07 to 0.59)	0.53 (0.24 to 0.81)
Spot the Change set size 5 (k)				
Effect size (95% CI)	−0.05 (−0.31 to 0.17)	0.17 (−0.05 to 0.47)	0.23 (−0.03 to 0.46)	0.40 (0.16 to 0.68)
Negative Emotion Recognition (number correct)				
Effect size (95% CI)	−0.02 (−0.28 to 0.21)	0.08 (−0.18 to 0.34)	0.27 (0.03 to 0.52)	0.49 (0.21 to 0.77)
UPSIT (number correct of 20)				
Effect size (95% CI)	−0.15 (−0.37 to 0.06)	−0.01 (−0.28 to 0.27)	0.28 (0.08 to 0.48)	0.38 (0.16 to 0.58)
Circle Tracing direct annulus length (log cm)				
Effect size (95% CI)	−0.01 (−0.25 to 0.26)	0.10 (−0.15 to 0.34)	0.15 (−0.12 to 0.40)	0.58 (0.33 to 0.85)
Circle Tracing indirect annulus length (log cm)				
Effect size (95% CI)	0.23 (−0.05 to 0.51)	0.19 (−0.10 to 0.48)	0.54 (0.30 to 0.79)	0.85 (0.58 to 1.18)

HD, Huntington's disease; ITI, inter-trial interval; SDMT, Symbol Digit Modalities Test; UPSIT, the University of Pennsylvania Smell Identification Test.

a premanifest sample of 117 individuals. This is important to note given that this premanifest sample had reasonably high levels of disease burden (mean=293.8), which yielded a median estimate of 10.8 years to onset. However, the sample also did not have significant motor signs indicative of HD at the time of study entry. This sample was designed to be a relatively pure sample which was unequivocally premanifest at the start of the study despite the disease burden scores indicating that they were in the latter premanifest stages. We anticipate that a premanifest sample that was closer to estimated disease onset or displaying significant motor signs could be expected to show greater degrees of cognitive change and that perhaps such changes would be detectable in a 24 month interval in a sample of about 120 participants. Indeed, we did find evidence for this in a partial examination of the cognitive battery within a smaller subsample of the TRACK-HD cohort.⁴ A test battery with a higher level of difficulty, designed specifically to challenge cognition in the premanifest period, might also be more likely to reveal decline over time. The Predict-HD cohort is also of great interest with regard to understanding the progression of cognitive change in the premanifest period in relation to disease burden and motor signs, and hopefully a longitudinal report of these data will be made available in the near future. Regardless, our findings suggest the plausibility of clinical trials for cognition in premanifest HD, and they highlight important issues for consideration of sample selection for such a trial.

This study makes several important contributions that will facilitate clinical trials to ameliorate cognitive decline in HD. First, to our knowledge, this is the only study of longitudinal cognitive assessment involving a battery of cognitive tests that has reported on both premanifest and HD groups, thus providing unique evidence of the relative sensitivity of tests to each other and across these stages of progression. Second, there

are few longitudinal reports in premanifest or diagnosed HD, and of these, none has as extensive a cognitive battery or as many participants or participant groups as TRACK-HD.^{11–15} Further, previous longitudinal cognitive studies used sample sizes too small to detect anything but large effects ($n < 25$), and/or batteries were strictly limited to one or to only a couple of cognitive tests. Finally, this study highlights the observation of differential practice effects across the groups as evidence of cognitive decline. Thus this report makes available, for the first time, a description of changes in cognition across a wide range of cognitive domains known to be affected in HD, across both premanifest and early HD, and across two annual follow-up time points.

Clinical trialists, because of the time restrictions they face in collection of data for clinical trials, must evaluate the relative sensitivity of outcome measures to select what they believe are the most sensitive measures. Provided that a putative treatment has the same proportionate effect on changes in all potential outcome measures (over and above the changes in healthy controls), outcome measures can be selected by comparing ES across measures. However, the fact that one ES is larger than another does not guarantee that the difference in the two ES is statistically significant, even if one ES is itself statistically significant and the other is not. For this reason, we coupled the construction of league tables of ES with pairwise comparisons to establish where there is evidence that particular ES are superior to others. Such an approach has significant benefits in the context of clinical trial planning because it provides an empirical basis with which to prioritise tests for inclusion in a clinical trial battery. The results showed us that there were no clear 'best' or 'worst' tests, and that instead, despite some differences in the magnitudes of the ES, many of the ES for the cognitive outcomes were not statistically significantly different from one

Table 6 Differences in standardised effect sizes of between group differences in longitudinal change over 12 and 24 months for pairs of variables adjusted for age, sex, centre and education

		Symbol digit		Circle tracing indirect		Stroop word reading		Circle tracing direct		Serial 2s with tapping		Paced tapping 3Hz		Negative emotion recognition		Spot the change set size 5		UPSIT smell identification		Paced tapping 1.8Hz		Trails A		Trails B		
		12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	12m	24m	
Circle tracing indirect	PreHD																									
	HD																									
Stroop word reading	PreHD																									
	HD																									
Circle tracing direct	PreHD																									
	HD	*	*	*	*																					
Serial 2s with tapping	PreHD																									
	HD	*	*																							
Paced tapping 3Hz	PreHD																									
	HD																									
Negative emotion recognition	PreHD																									
	HD	*	*																							
Spot the change set size 5	PreHD																									
	HD	*	*	*	*																					
UPSIT smell identification	PreHD	*	*	*	*																					
	HD	*	*	*	*	*	*																			
Paced tapping 1.8Hz	PreHD																									
	HD	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Trails A	PreHD																									
	HD	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Trails B	PreHD																									
	HD	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

The * symbol indicates a statistically significant difference (p<0.05) in the magnitude of the effect sizes (ES) for the pair of variables in question. For example, the ES for the difference in longitudinal change between early HD and controls on the Symbol Digit Modalities Test was found to be statistically significantly larger than the ES for Direct Circle Tracing over both 12 and 24 months. The lack of an * symbol in the corresponding cells for PreHD indicates that these ES for the differences in longitudinal change between the premanifest HD group and controls for this pair of variables were not statistically significantly different in magnitude at either 12 or 24 months. HD, Huntington's disease; UPSIT, the University of Pennsylvania Smell Identification Test.

another. For example, for both 12 and 24 month intervals, the SDMT had the largest ES in absolute terms. However, neither the 12 nor 24 month ES for SDMT was statistically significantly larger than the estimates for Stroop Word Reading, the indirect condition of Circle Tracing or the 3 Hz condition of Paced Tapping. Trails B had the smallest ES, but this ES was not significantly smaller than those for Negative Emotion Recognition, Spot the Change, Trails A or either of the Paced Tapping conditions.

Cognitive tests with the highest ES are likely to be the most statistically significant in a clinical trial of a disease modifying therapy, provided that such a therapy has a similar proportional effect on each test—that is, a drug that reduces the rate of decline in one test by 50% will also reduce the rate of decline in other tests by 50%. Of course, a more statistically significant effect does not necessarily translate into a more clinically significant effect but in the absence of information about which of the cognitive tasks considered here are most clinically important, this seems a reasonable criterion on which to base the selection of outcome variables for clinical trials.

A composite cognitive score may yield larger ES than those from individual cognitive tests but at present there is no well recognised cognitive combination that is used in practice. A number of statistical and non-statistical approaches could be used to derive such a score but there can be no certainty that a combination of cognitive outcomes with an increased ES will necessarily translate into a more statistically efficient

outcome for clinical trials. Specifically, if a treatment has non-proportional effects on the various test scores that make up a composite, then that composite may be less efficient than a composite score which emphasises the more responsive of the individual tests.

A clear understanding of where statistically significant differences in ES are and are not present also has implications for power analyses. For example, for the three tests with the largest ES at 12 months for early HD (SDMT, indirect Circle Tracing and Stroop Word Reading), sample size estimates for a 50% effective treatment, 90% power and two tailed p<0.05 group comparisons would be 150 (95% CI 75 to 374), 289 (95% CI 135 to 934) and 337 (95% CI 135 to 875), respectively, in each arm of a 1 year treatment trial with no dropouts. The results suggest that estimating sample sizes across the range of ES for equally best outcomes (in this case SDMT, indirect Circle Tracing and Stroop Word Reading) would help to avoid underestimating the sample needed. Given that ES are reduced by low reliability, and that cognitive outcomes tend to be relatively noisy measures, the findings also highlight the need to minimise noise wherever possible in measuring cognition. Thus careful control over standardised test administration and scoring is essential, as is minimisation of participant related variability linked to such factors as fatigue.

Due to the paucity of longitudinal studies, researchers must frequently utilise cross sectional results for selecting the most promising outcome measures. Yet cross sectional comparisons of

participants stratified along a continuum of progression may lead to gross overestimates of longitudinal effects over short time periods, such as those seen in clinical trials. For example, we previously reported cross sectional TRACK-HD findings⁵ for three cognitive outcomes that revealed significant group effects even though longitudinally we now show these measures to be among the least sensitive. Similarly, results from the very large cohort of premanifest HD participants in Predict-HD, which uses a set of cognitive measures that overlap with TRACK-HD, also show cross sectional sensitivity.¹⁶ Thus, wherever possible, ES for estimating samples sizes for clinical trials should be based on longitudinal observations, such as those reported in this paper.

Importantly, in diseases that affect cognition, ES estimates for rates of change conflate practice effects and deterioration. The possible impacts of this conflation must be carefully considered before using change rates to determine clinical trial sample sizes. In designing future studies or trials, attention should be given to using multiple baseline designs to help disentangle the contribution of practice to the observable changes from deterioration or treatment. ES from many tests also conflate motor deterioration and cognitive deterioration although the battery of tests we report here includes tests that can be argued specifically to be free of such confounds. Specifically, Spot the Change, Emotion Recognition and the UPSIT do not require rapid responding or precise movements, nor are their outcomes measured in terms of response speed. Thus deteriorations in performance in these tests, which were statistically significant in early HD, can be plausibly interpreted as indicating cognitive, but not motor, decline. When using the ES from cognitive batteries to determine power for clinical trials, it is important to keep in mind the potential interplay of cognition and motor function in order to select tests that are most suitable for the goals of a particular trial.

In conclusion, the findings from this study illustrate several considerations that are of general importance for designing cognitive outcome batteries for clinical trials, including the length of the follow-up needed, sensitivity of cognitive measures and the need to make careful assessments of whether ES are statistically different from each other. It also illustrates the limitations of using cross sectional findings to inform longitudinal designs.

Acknowledgements The authors offer their gratitude to the volunteers who participated and to their carers and companions who helped make their participation possible.

Contributors JCS was involved in obtaining funding, study design and conception, development and oversight of the cognitive battery, study set-up, data analysis and interpretation, and was responsible for drafting, revising and finalising the manuscript. She is the guarantor. RJ was involved in the analyses and interpretation of the data, as well as drafting, revising and finalising the manuscript. IL was involved in the development and oversight of the cognitive battery, design of the manuscript, interpretation of the data, and drafting, revising and finalising the manuscript. AO was involved in the development and oversight of the cognitive battery, monitoring of data collection, and revising and finalising the manuscript. SQ was involved in the development of the cognitive battery, interpretation of the data, and revising and finalising the manuscript. MJS, EMD, DJ, RDS, AC and EH were all involved in the preparation and set-up of the cognitive tools and materials at clinical sites, data collection, and revising and finalising the manuscript. AD, BRL and RACR were

involved in the concept and design of the study, and revising and finalising the manuscript. DRL was involved in study design and conception, protocol writing, statistical design and conception, data analysis and interpretation, and revising and finalising the manuscript. SJT is the study's global principal investigator, was involved in obtaining funding, study design and conception, and revising and finalising the manuscript. CF was involved in study design and conception, statistical design and conception, analyses and interpretation of the data, as well as drafting, revising and finalising the manuscript.

Funding TRACK-HD is supported by the CHDI/High Q Foundation Inc, a not for profit organisation dedicated to finding treatments for Huntington's disease.

Competing interests None.

Ethics approval Ethics approval was provided by University College London, Monash University, University of British Columbia, Leiden University Medical Centre and Hôpital de la Pitié-Salpêtrière.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement TRACK-HD is not an open access study, but CHDI and the study investigators are committed to ensuring that TRACK-HD data are used to define and validate the most promising endpoints for clinical trials in Huntington's disease using the best analytical, state of the art approaches available. The study includes 36 month longitudinal 3T MRI, clinical, cognitive, quantitative motor, oculomotor and neuropsychiatric measures, and a standardised plasma collection protocol. Requests for access to unpublished data should be sent to the study coordinators (coordination@track-hd.net) and will be considered by the Study Steering Committee on a case by case basis.

REFERENCES

1. **Gómez-Tortosa E**, MacDonald ME, Friend JC, *et al*. Quantitative neuropathological changes in presymptomatic Huntington's disease. *Ann Neurol* 2001;**49**:29–34.
2. **Gusella JF**, McNeil S, Persichetti F, *et al*. Huntington's disease. *Cold Spring Harb Symp Quant Biol* 1996;**61**:615–26.
3. **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 Neurol* 2011;**10**:31–42.
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 Neurol* 2012;**11**:42–53.
5. **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 Neurol* 2009;**8**:791–801.
6. **Penney JB**, Vonsattel JP, MacDonald ME, *et al*. CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann Neurol* 1997;**41**:689–92.
7. **Huntington study group**. Unified Huntington's disease rating Scale-99. Rochester, NY: Huntington Study Group, 1999.
8. **UNESCO**. *International Standard Classification of Education*. 1997. <http://www.uis.unesco.org> (accessed 16 Nov 2007).
9. **Efron B**, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, USA: Chapman & Hall, 1993.
10. **Altman DG**. *Practical Statistics for Medical Research*. London, UK: Chapman and Hall, 1991.
11. **Bachoud-Levi AC**, Maison P, Bartolomeo P, *et al*. Retest effects and cognitive decline in longitudinal follow-up of patients with early HD. *Neurology* 2001;**56**:1052–8.
12. **Beglinger LJ**, Duff K, Allison J, *et al*. Cognitive change in patients with Huntington disease on the repeatable battery of the assessment of neuropsychological status. *J Clin Exp Neuropsychol* 2010;**32**:573–8.
13. **Lemiere J**, Decruyenaere M, Evers-Kiebooms G, *et al*. Longitudinal study evaluating neuropsychological changes in so-called asymptomatic carriers of the Huntington's disease mutation after 1 year. *Acta Neurol Scand* 2002;**106**:131–41.
14. **Saleh N**, Moutereau S, Azulay JP, *et al*. High insulin-like growth factor I is associated with cognitive decline in Huntington disease. *Neurology* 2010;**75**:57–63.
15. **Witjes-Ane MN**, Mertens B, van Vugt JP, *et al*. Longitudinal evaluation of "presymptomatic" carriers of Huntington's disease. *J Neuropsychiatry Clin Neurosci* 2007;**19**:310–17.
16. **Stout JC**, Paulsen J, Queller S, *et al*. Neurocognitive signs in prodromal Huntington disease. *Neuropsychology* 2011;**25**:1–14.