Activity- is better than connectivity-neurofeedback training in Huntington's disease

Marina Papoutsi* (1), Joerg Magerkurth (2), Oliver Josephs (3), Sophia E Pépés (4), Temi Ibitoye (1), Ralf Reilmann (5, 6), Nigel Hunt (7), Edwin Payne (7), Nikolaus Weiskopf (3, 8), Douglas Langbehn (9), Geraint Rees† (3, 10), Sarah J Tabrizi† (1)

(1) UCL Huntington's Disease Centre, Queen Square Institute of Neurology, University College London, UK, (2) Birkbeck-UCL Centre for Neuroimaging, University College London, London, UK, (3) Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, UK, (4) University of Oxford, UK, (5) George Huntington Institute and Dept. of Radiology University of Muenster, Germany, (6) Section for Neurodegeneration and Hertie Institute for Clinical Brain Research, University of Tuebingen, Germany, (7) Eastman Dental Institute, University College London, UK, (8) Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, (9) Carver College of Medicine, University of Iowa, USA, (10) Institute of Cognitive Neuroscience, University College London, UK

† Equal senior authors

* Corresponding author:

Dr Marina Papoutsi

Email: m.papoutsi@ucl.ac.uk

Address: UCL Huntington's disease centre,

Russell Square House,

10-12 Russell Square,

London, WC1B 5EH, UK

Abstract

Neurofeedback training (NFT) could support cognitive symptom management in neurodegenerative diseases such as Huntington's disease (HD) by targeting brain regions whose function is disrupted by the disease. Identifying the most appropriate target for NFT is not straightforward. The aim of our study was to test whether HD patients can learn to regulate their brain activity using NFT and to compare two different NFT targets, activity NFT using as target the activity from the Supplementary Motor Area (SMA) and connectivity NFT using the correlation between SMA and left striatum signal. To evaluate each approach we measured learning by testing for an increase in NFT target levels across training visits, and near transfer by examining upregulation of the target levels in the absence of feedback after training. The activity NFT treatment group was the only group that showed both successful learning and near transfer, suggesting that it's a more promising approach in HD than connectivity NFT.

Introduction

Neurofeedback training (NFT) is a non-invasive intervention used to train participants in a closed-loop design to regulate their own brain activity[1]. The underlying principle is that by regulating different aspects of their brain activity, e.g. regional activation or inter-regional connectivity, participants would implicitly regulate associated cognitive function. Because NFT can be delivered non-invasively, it can be used in clinical populations either preventatively or as an adjunct treatment to other potential disease-modifying therapies. However, there are several challenges in designing NFT trials, including the choice of an appropriate NFT target for the specified clinical population.

Huntington's disease (HD) is a genetic neurodegenerative condition characterised by progressive motor, psychiatric and cognitive impairment, as well as early striatal atrophy, cortical and cortico-striatal connectivity loss[2–5]. Striatal activity and cortico-striatal connectivity would therefore be obvious targets for NFT. However, because of the atrophy present in the striatum in HD[5,6], as well as the increase in iron deposits[7], BOLD fMRI signal from the striatum may be difficult to measure reliably in real-time in HD patients. Therefore signal from cortical regions that connect to the striatum, and can be measured reliably might be more appropriate. Previous studies have shown that NFT induced changes are not just localised to the target region, but extend to a wider network of regions[8–10], suggesting that a proxy region might also be appropriate.

In a recent proof-of-concept study we used the supplementary motor area (SMA) as a target for real-time fMRI NFT in HD patients[11]. We selected BOLD fMRI signal from the SMA because it can be reliably measured in real-time[12,13], and its function and connectivity to the striatum is disrupted by HD[14]. We showed that HD patients can be trained to increase the level of SMA activity and that improvement in cognitive and psychomotor behaviour after training related to increases in activity of the left Putamen and SMA – left Putamen connectivity during training. This suggested that SMA-striatum connectivity could be a more appropriate NFT target than SMA activity in HD.

The aim of the current study was therefore to compare the two NFT approaches in order to determine which one is better, and to collect further evidence on the feasibility of the method in HD. For this purpose we used BOLD fMRI signal change from the SMA as the target for activity NFT and correlation between signal from the SMA and left striatum during upregulation as the target for connectivity NFT[15,16]. Connectivity NFT feedback was presented intermittently at the end of the upregulation block, whereas during activity NFT feedback was provided continuously during upregulation, similar to our earlier work. Because of these differences, we would not be directly

comparing the two groups, but rather we will examine changes within each group separately, and also compared them to matched control groups that received non-contingent, sham neurofeedback. Participants that were randomized to the control group were yoked to a participant in the treatment group, and received feedback based on the activity of their yoked participant from the treatment group, rather than their own. This ensured blinding of all participants and controlled for experimental exposure and motivation. Comparisons between the treatment and control groups would enable us to estimate effect sizes (ES) that would then be used to inform future larger trials testing for efficacy of NFT in HD.

Results

To identify which approach is better we examined learning and transfer success in the two different NFT types. We tested two main effects: 1) training effect[17], defined as a linear increase in the target NFT measures from the baseline to the last training session, and 2) transfer effect[17], defined as an increase in the follow-up (after training) sessions compared to baseline in both imaging and behavioural measures. Because the two NFT approaches use a different feedback measure, contrast estimates in the case of activity NFT and correlation coefficients in the case of connectivity NFT, we could not compare the two types of NFT directly. Instead we performed within group analyses and compared each treatment group with its matched control group and reported effect sizes for each comparison.

Baseline Levels of the NFT Target Measure

The baseline session included a motor imagery task used to calculate the levels of the NFT target measure prior to NFT (see Methods section for details on experimental design). For the activity NFT group this was the BOLD fMRI signal from the SMA during motor imagery minus the baseline SMA activity, whereas for the connectivity NFT group it was the time-series correlation between the SMA and the left striatum during upregulation (SMA-striatum connectivity). We first examined whether there were any differences between the groups at baseline. A one-way ANOVA with group as the main effect (treatment vs control) showed that for the activity NFT group the main effect of group was not significant ($F(1, 14) = 0.174$, $p = 0.683$, estimate(SE) = -0.089(0.214)). However, the intercept was ($t(14) = 5.916$, $p < 0.001$, estimate(SE) = 0.894(0.151)), suggesting that participants were able to reliably increase SMA activity prior to NFT. For the connectivity NFT group neither the main effect of

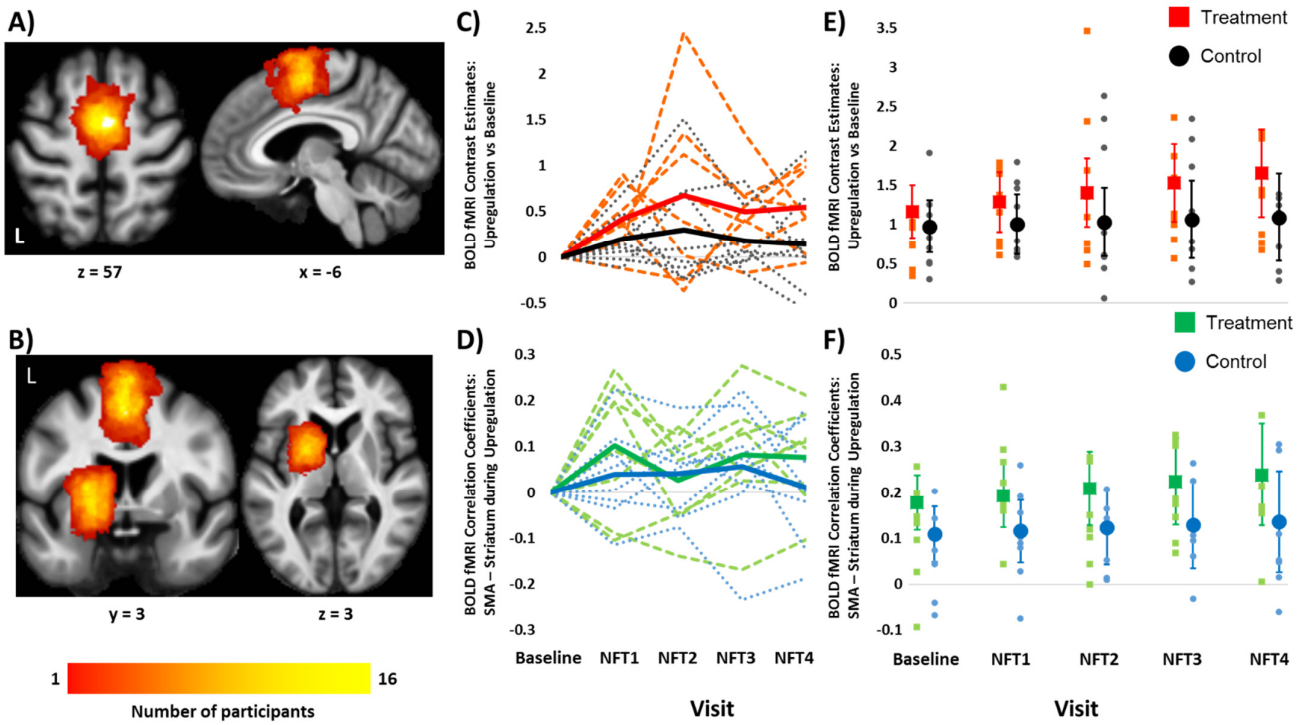

Figure 1: Learning effects. (A) and (B) Heat maps showing the location and overlap of ROI voxels across all participants in the activity and connectivity NFT groups respectively. Maps are superimposed on a group average MT image. (C) and (D) Change from baseline in the target NFT levels across all training sessions per subject (dotted lines). The group mean per session is shown with thick continuous lines. Shown in red and green hues are the treatment groups, whereas shown in black and blue hues are the control groups for the activity (C) and connectivity (D) NFT groups respectively. The scaling of the plots is different between plots C and D and therefore they are not visually comparable. (E) and (F) Dot plots show the NFT target levels across visits for activity (E) and connectivity (F) NFT respectively. Squares represent the treatment groups, whereas circles represent the control groups (red and green for the treatment groups; black and blue for the control groups). The small squares and circles show the individual data points, whereas the larger squares and circles show the adjusted mean group effects. Error bars are 95% CI.

group was significant ($F(1, 14) = 1.171$, $p = 0.298$, estimate(SE) = 0.054(0.050)), nor the intercept ($t(14) = 1.793$, $p = 0.095$, estimate(SE) = 0.064(0.035)). This suggested that SMA-striatum connectivity was not reliably engaged prior to NFT.

Learning Effects: Linear increase across sessions

To examine learning across the two different types of NFT we first tested for a linear increase in the target NFT measure within each group using linear mixed models. The dependent variables were NFT target ROI contrast estimates in the case of the activity NFT group and correlation coefficients in the case of the connectivity NFT group. Figure 1A and 1B show the location of the selected ROIs. Session was included as a repeated fixed effects factor and modelled as a continuous variable with values increasing linearly from the baseline to the last training session (i.e. 1 to 5).

For the activity NFT treatment group there was a significant effect of session, whereby activity in the target ROI increased from the baseline to the last training session ($F(1, 14.4) = 11.31$, $p = 0.005$, slope(SE) = 0.128(0.038). In contrast, neither the connectivity NFT treatment group nor any the control groups showed a significant effect of session (connectivity NFT treatment $F(1, 17.9) = 4.01$, $p = 0.06$, slope(SE) = 0.017(0.009); activity NFT control ($F(1, 19.1) = 0.61$; $p = 0.446$, slope(SE) = 0.024(0.031)); and connectivity NFT control $F(1, 25.6) = 1.07$, $p = 0.311$, slope(SE) = 0.010(0.009)). Figure 1C and 1D show change in NFT target levels from baseline per participant (dashed lines), as well as the group mean (thick line), for activity and connectivity NFT respectively.

To test for differences between treatment and control groups within activity NFT or connectivity NFT we ran the same model as above adding group as a fixed effect factor. The group by session interaction in the model shows if there were any differences in the learning slopes between the treatment and control groups. For activity NFT there was a significant main effect of session ($F(1, 33.8) = 8.89$, $p = 0.005$), but not a significant main effect of group ($F(1, 20.8) = 0.21$, $p = 0.650$), or a group by session interaction ($F(1, 33.8) = 3.38$, $p = 0.075$, estimate(SE) = 0.093 (0.051), effect size = 0.65). The results suggest there was a difference across sessions in both treatment and control groups, with baseline being lower than the training sessions across both group (Figure 1E). For connectivity NFT there was no significant main effect of session ($F(1, 43.6) = 2.82$, $p = 0.100$), group ($F(1, 15.2) = 2.55$, $p = 0.131$), or group by session interaction ($F(1, 43.6) = 0.43$, $p = 0.517$, estimate(SE) = 0.009 (0.013), effect size = 0.23; Figure 1F).

Near Transfer: Upregulation without feedback

To measure whether participants were able to increase the NFT target levels volitionally after training and without feedback (known as near transfer), we examined change from baseline in the follow-up sessions. There were three follow-up sessions, within 2 weeks of training, between 4-6 weeks and between 8-10 weeks (see Methods for more details). At the follow-up sessions participants performed the same task as at baseline, but were instructed to use the mental strategy that they believed was the most effective in upregulating the NFT target during the training sessions. To examine within-group changes we used linear regression with dependent variable the change in the target NFT measure from the baseline to each of the follow-up visits and adjusting for the baseline level of the target NFT. Adjusting for baseline levels is used to increase sensitivity in pre-post treatment comparisons[18]. In these analyses the intercept was the effect of interest and represented adjusted change at follow-up from baseline.

At the first follow-up the intercept in the treatment activity NFT group was significant ($t(6) = 3.03$, $p = 0.023$, estimate(SE) = 0.386(0.127)), suggesting that the group significantly increased the target NFT activity compared to baseline. Similarly, the intercept was significant in the treatment connectivity NFT group ($t(6) = 2.89$, $p = 0.028$, estimate(SE)=0.123(0.043)). In contrast, the intercept was not significant for the activity NFT control group ($t(6) = 1.02$, $p = 0.346$, estimate(SE) = 0.179(0.176)), or for the connectivity NFT control group ($t(5) = 1.13$, $p = 0.311$, estimate(SE) = 0.051(0.045)). In summary, only the two treatment groups showed a significant change from baseline at the first follow-up.

Subsequent analyses of the second and third follow-up sessions revealed that differences between the follow-up and the baseline sessions were not reliably sustained for either the treatment, or the control groups. The activity NFT treatment group, which showed significant changes during the first follow-up compared to baseline, only showed significant changes during the third follow-up ($t(6) = 2.49$, $p = 0.047$, estimate(SE) = 0.303(0.122)), but not the second ($t(5) = 0.54$, $p = 0.616$, estimate(SE) = 0.102(0.190); Figure 2A). None of the other groups showed any significant changes for either the second or the third follow-up (all $p > 0.07$; see Supplementary Table 1).

To test between-group differences we used an ANCOVA to compare change from baseline to the first follow-up in the treatment and control groups, controlling for baseline NFT target levels. Because we used the change scores in the ANCOVA, the group main effect was the effect of interest and equivalent to a group by session interaction looking at the difference between groups in the change from baseline. There were no significant main effects of group for either the activity NFT ($F(1, 13) = 0.54$, $p = 0.475$, estimate(SE) = 0.186(0.253), effect size = 0.37; Figure 2), or the connectivity NFT ($F(1, 12) = 3.07$, $p = 0.105$, estimate (SE) = 0.124(0.071), effect size = 0.98; Figure 2B). Similarly, neither the second nor the third follow-up sessions showed significant change compared to baseline between the treatment and control groups (all $p > 0.2$; see Figure 2 and Supplementary Table 2).

Far Transfer: Cognitive and psychomotor performance

To assess the effects that NFT had on the participants' performance in tasks unrelated to the training (far transfer), we compared performance in a number of cognitive and psychomotor tasks selected a priori because they have been previously shown to be sensitive to HD progression[5,11] before and after training. They were then standardized and summed to create a behavioural composite score (see Methods section for details). The change in the individual scores is shown in Supplementary Figure 1 for completion. Behavioural performance was measured twice prior to the start of NFT and

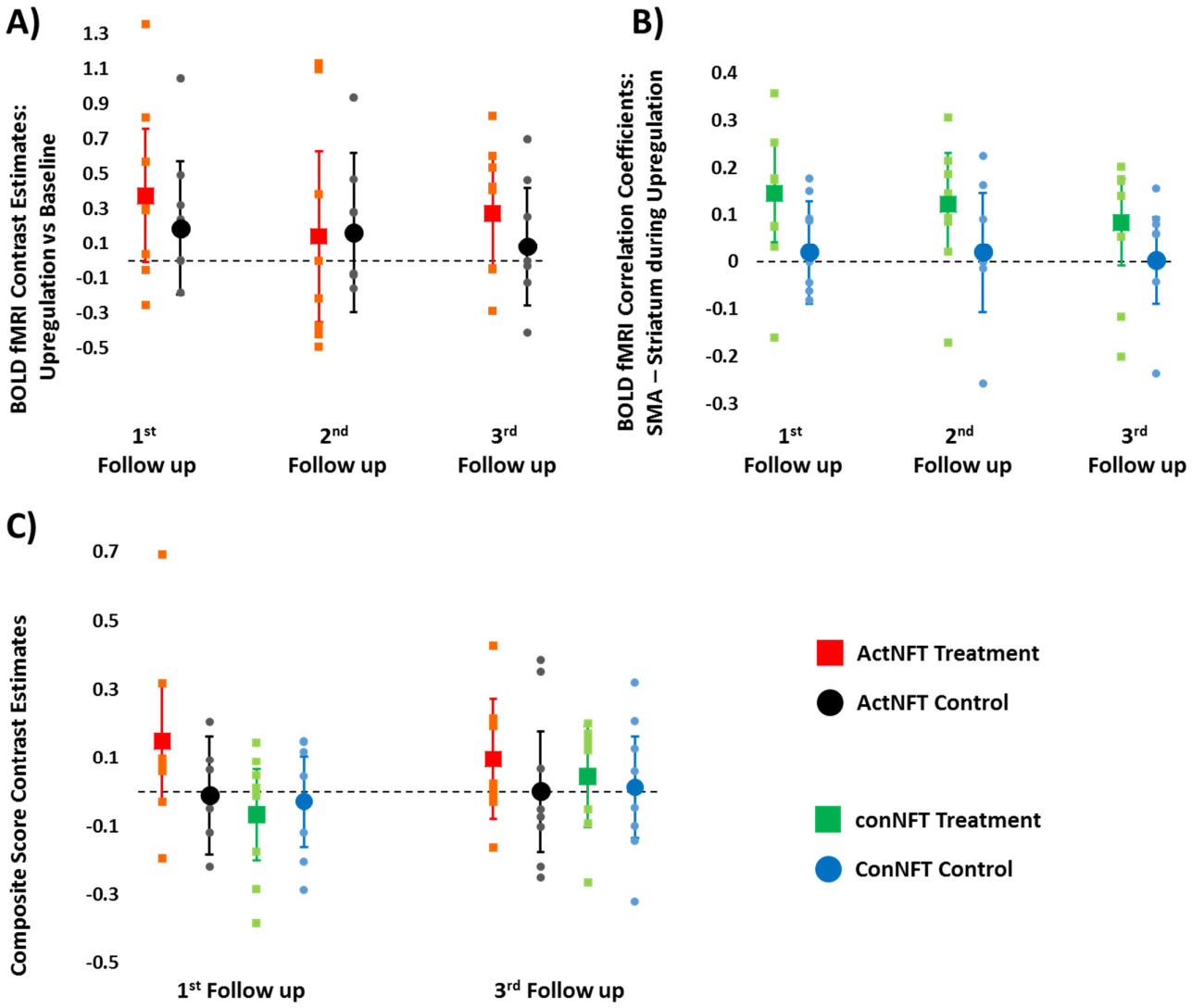

Figure 2: Near and far transfer effects. (A) and (B) Dot plots show the change in NFT target levels compared to baseline across follow-up visits for activity and connectivity NFT respectively. (C) Dot plots show the change in the behavioural composite score compared to baseline for all groups. The horizontal black dashed lines in all three plots show the baseline, data points above this line represent an increase compared to baseline. Squares represent the treatment groups, whereas circles represent the control groups (red and green for the treatment groups; black and blue for the control groups). The small squares and circles show the individual data points, whereas the larger squares and circles show the adjusted mean group effects. Error bars are 95% CI.

the second session was used as baseline to account for practice effects. It was also measured twice after the end of NFT, within two weeks and between 8-10 weeks after the last training session. Although the study was not powered to detect changes in behaviour, estimating the effect size would be useful when planning future research studies that would be powered on behavioural change, therefore it was important to include behavioural measurements.

None of the groups showed significant increase in the composite score in either of the follow-up sessions compared to baseline (all $p > 0.19$; see Supplementary Table 3). Overall, in the activity NFT

treatment group, 6 out of 8 participants (75%) had better performance at the first follow-up compared to baseline, and 5 out of 8 participants (62.5%) had better performance at the second follow-up compared to baseline (mean(SE) change compared to baseline at first follow-up = 0.145 (0.100) and at second follow-up = 0.084 (0.070)). In the activity NFT control group, 4 out of 8 (50%) had better performance at the first follow-up compared to baseline, and 3 out of 8 (37.5%) at the second-follow up (mean (SD) change compared to baseline at first follow-up = -0.003 (0.050) and at second follow-up = 0.014 (0.089)). Similarly, in the connectivity NFT treatment group, 4 out of 8 (50%) had better performance at the first follow-up compared to baseline, and 5 out of 8 (62.5%) at the second-follow up (mean (SD) change compared to baseline at first follow-up = -0.069 (0.051) and at second follow-up = 0.046 (0.063)). Lastly, in the connectivity NFT control group 4 out of 8 (50%) had better performance at the first follow-up compared to baseline and at the second-follow up (mean (SD) change compared to baseline at first follow-up = -0.023 (0.063) and at second follow-up = 0.014 (0.079)).

To compare change in performance across groups we used ANCOVAs with fixed effects of group (treatment vs control) and adjusting for baseline performance. As previously we used change from baseline as our dependent measures, therefore the main effect of group was equivalent to a group by session interaction. For activity NFT there were no significant group differences for either the first ($F(1, 13) = 1.86$, $p = 0.195$, estimate(SE) = 0.161(0.118), effect size = 0.74), or the last follow-up ($F(1, 13) = 0.62$, $p = 0.444$, estimate(SE) = 0.095(0.121), effect size = 0.43). This was similar for connectivity NFT, there were no significant group differences for either the first ($F(1, 13) = 0.20$, $p = 0.665$, estimate(SE) = -0.039(0.088), effect size = -0.22) or the last follow-up ($F(1, 13) = 0.10$, $p = 0.756$, estimate(SE) = 0.031 (0.097), effect size = 0.16).

Power Calculations

Based on the results presented so far only the activity NFT treatment group showed a significant increase in NFT target levels during training and successful upregulation of NFT target levels in the absence of training (near transfer). Although the connectivity NFT treatment group showed upregulation of the NFT levels in the absence of training, there was no significant linear increase in the NFT levels across the training visits. In addition, although none of the groups showed a statistically significant within group change during the first follow-up visit in the behavioural composite score compared to baseline or a significant group by session interaction, only the activity NFT group had a positive estimate of change in both within- and between-group comparisons. Our

results therefore suggest that the activity NFT treatment group is a more plausible NFT type for HD and more appropriate as a target in future trials.

A future RCT using NFT designed to show efficacy would need to be powered on the basis of far transfer effects on a behavioural endpoint. To calculate the effect size (Cohen's d) for a power calculation we ran a two-sample t-test assuming equal variances to compare baseline to first follow-up change in the behavioural composite score between the activity NFT treatment vs control groups ($t(14) = 1.42$, $p = 0.178$, effect size = 0.71). To detect an effect size of 0.71 in this context with a type I error rate of 5% and 80% statistical power, we would need 33 subjects per group (assuming no follow-up loss).

Discussion

The aim of the present study was to compare two different NFT methods, activity and connectivity NFT, and establish which one is preferable to use in HD. We replicated our previous findings that HD patients can learn to increase their target NFT levels during training using SMA activity as NFT target and further showed that activity NFT is more promising than connectivity NFT. The activity NFT treatment group was the only group that showed successful learning and near transfer, as well as more promise in terms of behavioural change after NFT. Although the connectivity NFT treatment group showed successful near transfer, it did not show successful learning or far transfer; if anything the results from the behavioural performance were on the opposite direction, suggesting that it could be an unfavourable approach. The results from our study, combined with the fact that SMA activity NFT is much simpler to administer and setup, led us to conclude that SMA activity NFT is more preferable than SMA-striatum connectivity NFT, in the case of our patient population.

A fundamental difference between the two NFT approaches is the type of signal provided as feedback during training. In the case of activity NFT we presented participants in the treatment group with percent change in SMA activity during upregulation vs baseline. In the case of connectivity NFT we presented participants in the treatment group with the correlation coefficients between signals recorded from the SMA and from the left striatum during upregulation. In the first case, percent signal change can be computed and presented as feedback to participants continuously in near real-time, which means that there is greater perceived contingency between a participant's mental actions and the feedback they receive. In the second case, correlations are computed over a number of time-points, in our case 30s, and the feedback is presented intermittently at the end of the upregulation block, which results in lower contingency. In our case the two elements, frequency of feedback presentation and NFT type, are intertwined and it is

therefore not possible to identify whether the differences observed are driven by the lack of contingency or the type of NFT target.

A previous study comparing continuous vs intermittent feedback using percent signal change in the amygdala in healthy young adults showed that participants were able to learn to increase the target NFT levels using both approaches, although intermittent feedback was more effective than continuous in that study[17]. These results suggest that the differences observed in our study could be driven by the type of feedback presented and not the frequency of presentation. However, both activity[17,19] and connectivity[15,16,20] NFT have been used successfully in other studies, suggesting that both methods could be appropriate. A possible explanation for our results could be that the quality of the signal recorded from the striatum in real-time was not reliable. Atrophy in the striatum is one of the earliest measurable signs of HD pathology in the brain and is accompanied by a decrease in striatal BOLD fMRI activation[21,22]. This would have an effect on the local signal-to-noise ratio (SNR) and our ability to measure striatal activation reliably. For the purposes of this study we defined the striatal ROI functionally and used a large enough region to try to mitigate this issue by averaging across functionally relevant voxels. Even so, the SNR for the SMA will be higher, therefore the activity NFT group would have received more accurate real-time feedback than the connectivity NFT group, which could explain the differences in learning between the two groups.

Because of the differences between the two NFT types, it was not possible to directly compare the two different treatment groups. Instead, we compared each treatment group to a matched sham control group. We chose a sham control group in order to control for potential placebo effects as a result of recruiting participants to an interventional study[23]. By choosing the "yoked" approach we ensured that the feedback control participants received was biologically plausible and matched to that of the treatment group. We chose not to use the approach of using a different ROI for the control group, because of potential problems with the spread of training effects across other brain regions. We do not yet understand the mechanism underlying NFT in HD and how widespread any effects could be, therefore we were not certain which other regions in the brain would be appropriate to be a control target[24]. Subsequent analyses examining the relationship between the signals of the yoked pairs (see supplementary methods on sham neurofeedback) show that our approach worked. The correlation between the signal from the control participants' brain and the signal from the brain of the yoked participants was overall quite low, suggesting that the feedback the control group received was not contingent on their actual brain activity.

A limitation of the present study was that it was only single-blind. This means that although participants were not aware of their group allocation, the researchers conducting the MRI and

behavioural assessment were not blinded. Because this was a small, feasibility study, double-blinding was not possible at this stage. To minimize any researcher bias, patients were not provided any input during the NFT sessions, but relied on the feedback they received during NFT. Statistical analyses were pre-defined and all behavioural assessments were objective measures of computer- or paper-based tasks. In addition, the main aim of this study was not to test the efficacy of the approach, but rather to compare two different NFT methods. None of the two methods was deemed more preferable at the start of the study, therefore researcher bias was minimal, if any. Double-blinding would, however, be necessary for any future RCTs that would test efficacy of the method.

Another limitation of the study was the small sample size. Because this is the first sham-controlled neurofeedback study in HD, we based our sample size calculations on a previous proof of concept NFT study on healthy young adults, which showed large effect size with group size of 11 per group[25] (see Methods section on participants). However, the reported effect sizes in healthy controls may not be appropriate for a study in clinical populations. Hence an additional aim of this study was to calculate effect sizes appropriate for our population and experimental design. The observed effect size for far transfer (behavioural change) at the first follow-up was moderate and sample size calculations suggest that a future RCT would need at least 33 participants per group in order to have 80% statistical power. This is a feasible number of participants to recruit in an RCT and further highlights that NFT using SMA activity may be a promising non-invasive intervention for HD.

To conclude, the aim of our study was to compare two different NFT approaches in HD, SMA activity and SMA – left striatum connectivity NFT. Although cortico-striatal connectivity is biologically more relevant in HD, the results from our study suggest that SMA activity NFT is a more plausible approach than connectivity NFT. SMA activity NFT is simpler to administer and the moderate effect sizes calculated in our study suggest that this approach may hold promise, as a non-invasive preventative or adjunctive intervention in HD. A future larger RCT is required in order to collect more robust evidence on the efficacy of the approach for the treatment of cognitive and psychomotor symptoms in HD for which there are currently no available treatments.

Materials and Methods

Participants

Thirty-four adults who carried an HTT gene CAG expansion greater than 40 were recruited to the study. One participant withdrew from the study after three visits because he could not tolerate the

		Activity NFT		Connectivity NFT	
		Treatment Group	**Control Group**	**Treatment Group**	**Control Group**
Number of Participants		8	8	8	8
Gender		6F, 2M	6F, 2M	6F, 2M	5F, 3M
Handedness		7RH, 1LH	7RH, 1LH	7RH, 1LH	8RH, 0LH
Age: Mean (SD)		46.4 (11.3)	50 (12.3)	52.3 (11.9)	50.1 (10.3)
CAG Repeat Length: Median (SD)		43 (3.7)	42.5 (2.1)	43 (2.5)	43.5 (1.4)
CAP Score: Mean (SD)		92.7 (14.2)	97.6 (11.7)	105.6 (23)	101.9 (18.3)
UHDRS	***TMS: Mean (SD)***	8 (12.7)	8.5 (4.3)	9 (10.1)	11.5 (14.1)
	TFC: Mean (SD)	11.6 (1.5)	12.5 (1.1)	12.5 (0.5)	11.6 (1.9)
MoCA: Mean (SD)		26.1 (4.2)	27.6 (1.1)	25.4 (3.3)	25.5 (3.0)
HADS	***Anxiety: Mean (SD)***	4.0 (2.3)	3.5 (3.9)	4.3 (3.3)	4.6 (4.9)
	Depression: Mean (SD)	1.8 (0.7)	1.9 (1.9)	3.6 (4.8)	3.6 (3.0)
Composite Score: Mean (SD)		-0.52 (0.75)	-0.21 (0.37)	-0.82 (0.67)	-0.95 (1.33)

Table 1: Demographic Information. Group differences between treatment and control groups were tested using a Mann-Whitney non-parametric test. All differences were non-significant (all p > 0.2). NFT: Neurofeedback training. CAP: normalized CAG-Age Product Score.

MRI scanning environment; the data were not used in any of the analyses. Another participant was excluded from the study, because a large number of trials had to be excluded due to motion related artifacts (see section on offline data analysis below for more details). These issues were identified during data pre-processing and another participant was recruited as a substitute. The remaining thirty-two participants who completed the training and testing protocol were included in the analyses (23 females, mean (SD) age = 49.7 (11.1)). There were no statistically significant differences between the treatment and control groups for the two types of NFT for any of demographic measures (using a non-parametric Mann-Whitney test all $p > 0.2$; Table 1 for detailed participant information). All participants provided written informed consent according to the Declaration of Helsinki and the study was approved by the Queen Square Research Ethics Committee (05Q051274). All procedures, including recruitment, consent and testing, were carried out in accordance to the relevant good clinical practice guidelines and regulations. Information regarding power sample calculations prior to the start of the study are provided in the supplementary methods.

Study Structure

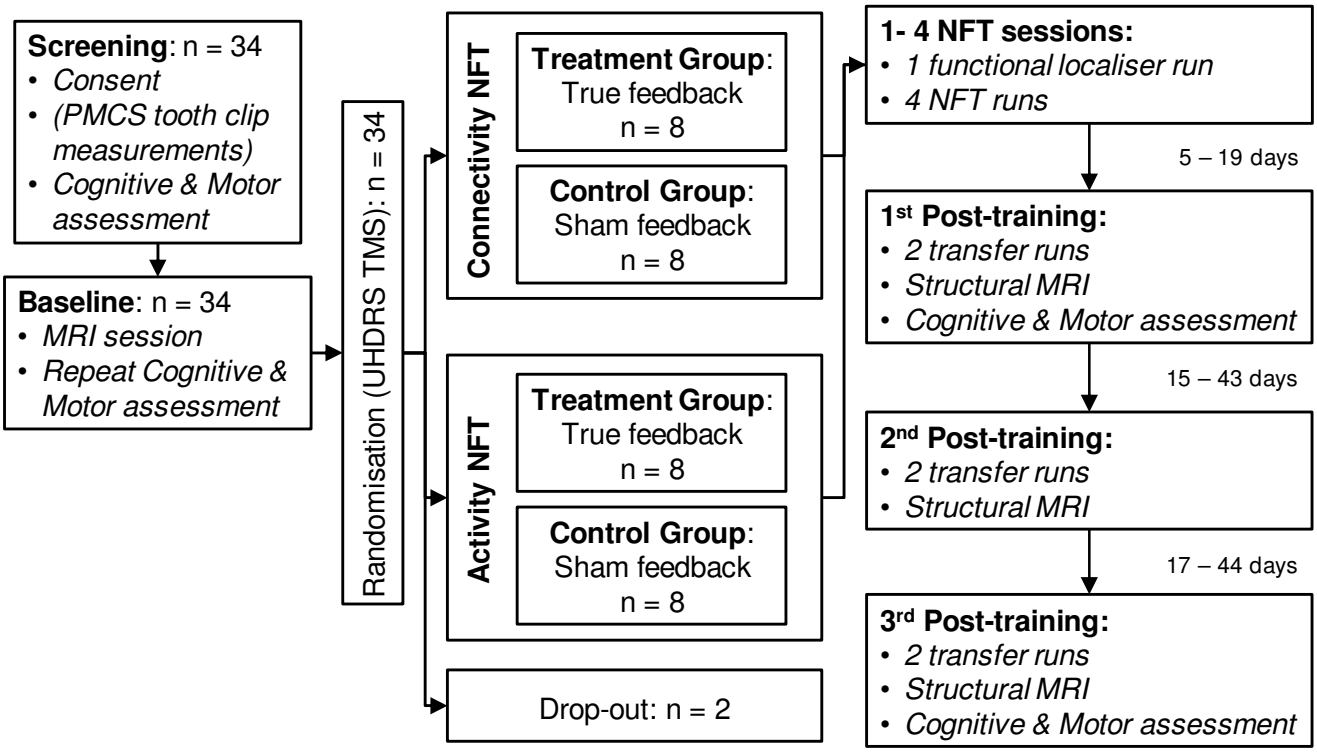

Figure 3: Diagram of study structure

As part of the study, participants completed 1 screening, 1 baseline, 4 neurofeedback training and 3 follow-up sessions. A diagram of the study design is shown in Figure 3. Supplementary Table 4 shows participant retention across visits and groups. After the baseline session participants were randomized to one of four groups: activity NFT treatment and control groups, and connectivity NFT treatment and control groups. Randomization was based on the Unified Huntington's Disease Rating Scale[26] (UHDRS) Total Motor Score (TMS). More details regarding the randomization procedure are provided in the supplementary methods.

Screening Session

During the screening session participants provided consent, completed the Montreal Cognitive Assessment test (MoCA)[27] of cognitive capacity, and performed a number of cognitive and psychomotor tasks (see section below). The same measurements were repeated during the baseline and follow-up sessions. The purpose of the testing during the screening visit was to familiarize the participants with the tests and minimize practice effects. Participants who consented to the use of a Prospective Motion Correction system (PMCS) to correct head motion at scan acquisition also had their teeth impression taken during that visit by a qualified orthodontist (NH). Details of the PMCS are provided in the supplementary methods.

Baseline & Follow-up Sessions

There was one baseline session and three follow-up sessions. The first follow-up was within 2 weeks from the last training visit, the second between 4-6 weeks and the third between 8 and 10 weeks (Supplementary Table 4). The baseline and follow-up sessions included: 1) repetition of the cognitive and psychomotor testing (only on the first and third follow-up), 2) structural MRI measurements using multi-parameter maps (MPMs)[28]and 3) two fMRI runs assessing the participant's ability to upregulate their motor control network.

At the baseline session participants were instructed to use motor imagery during the active blocks in order to increase activity in regions of their brain involved in movement and motor control. The aim was to measure the NFT targets' baseline activity/connectivity levels prior to training. At the follow-up sessions participants were instructed to use the mental strategy that they believed worked best to increase the level of the NFT targets during the training. These "near transfer" runs measure whether participants were able to regulate the NFT target in the absence of feedback and therefore assess learning. It is important to note that in this study we did not explicitly ask participants to practice upregulation at home between the end of the training and the follow-up sessions in order to measure how long any effects of training can be sustained without any additional practice.

The fMRI runs consisted of 5 upregulation blocks (30s each), 6 rest blocks (30s each) and 5 response blocks (18s each). Similar to our previous study[11] we used a simple attention task during the rest blocks, whereby participants monitored changes in the luminance of a white bar. If the white bar flickered to grey, they would wait until a question mark appeared inside the white bar (response blocks) and then make a response by clenching their left fist once. A maximum 3 out of 6 baseline blocks flickered per run and the timing of the flickering during the block was random. The design of the tasks is shown in Figure 4B.

Neurofeedback Training Sessions

All NFT sessions started with a functional localiser run to identify the target ROIs. Participants were instructed to clench their left fist during the active blocks (10 blocks lasting 20.4s each) and rest during the rest blocks (11 blocks lasting 20.4s each). The design of the run is shown in Figure 4A. Using Turbo-BrainVoyager (TBV; Brain Innovation, The Netherlands) the fMRI run was analysed in real-time and the resulting statistical map was used to define the target ROIs for the subsequent NFT

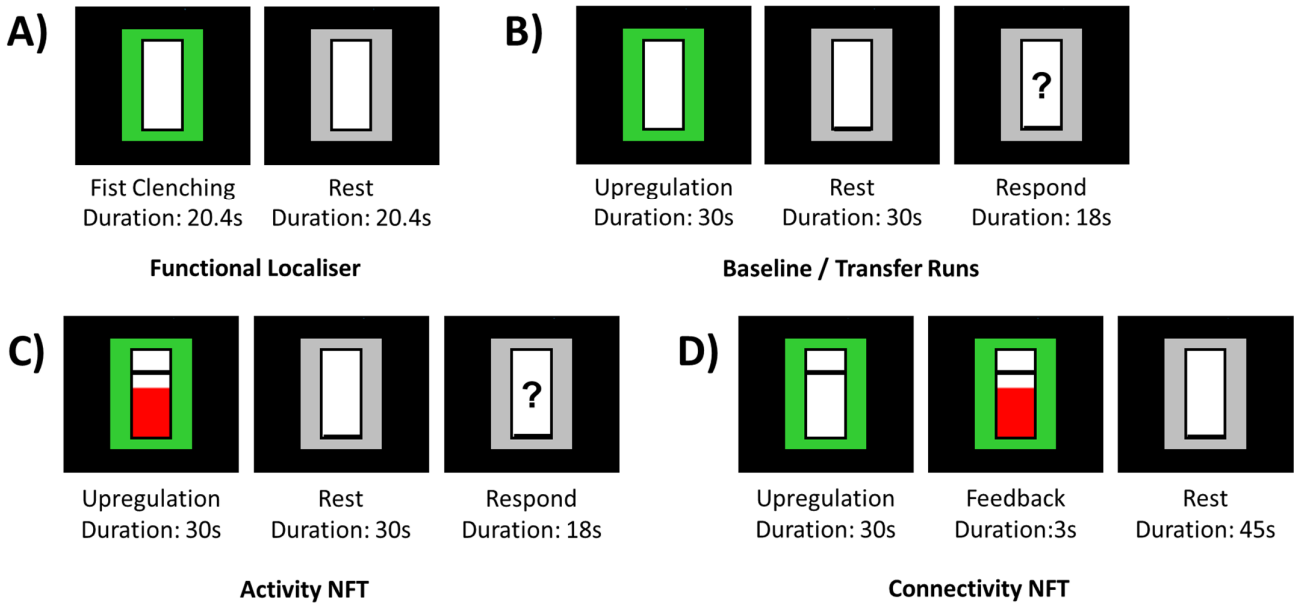

Figure 4: Task design. (A) – (D) All tasks consisted of alternating blocks. The figure shows the screen that the participants viewed. The setup of the runs was consistent across all different sessions to condition the participants such that a green frame was associated with action, whereas a gray frame was associated with rest.

runs. For the activity NFT sessions, the SMA was selected as the target ROI. For the connectivity NFT sessions, the SMA and the left striatum (including putamen, globus pallidus and caudate) were selected as the target ROIs. The ROIs were drawn using TBV. It was not always possible to acquire a structural MRI volume during the baseline visit, because of fatigue. To avoid having to add an extra visit and burden the patients, we added the structural scan at the start of their first neurofeedback training session, and chose to define our ROIs using a functional localiser and anatomical landmarks, rather than creating an anatomical mask of the region. For the SMA the statistical map was thresholded at t-value = 3 and a rectangle was drawn around the SMA cluster for the active vs rest contrast. The location of the striatum was identified visually on the first EPI scan of the localiser run using landmarks and the EPI contrast. Due to high iron concentration, the putamen and globus pallidus appear darker on an EPI scan and are therefore easy to identify on EPI scans. A rectangle was drawn around the striatum including the putamen, globus pallidus, caudate and ventral striatum. Because of the rectangular shape, the striatal ROI, also included surrounding white matter. However, the ROI was centred around the striatum and most of the recorded signal originated from the gray matter of the striatum. To enhance SNR we further applied 6mm FWHM smoothing during online data acquisition and also corrected in real-time for head motion and physiological noise (see below for more details).

Similar to our previous study[11] and comparable to other studies[12,29–31], the ROIs were re-drawn at each session ensuring that only voxels with high activation are selected. The ROIs from the first visit

were used as a reference, when drawing the ROIs for the subsequent visits to ensure that the position was similar, although the exact voxels selected might be different. A heat map showing the overlap of the ROIs across all participants is shown in Figure 1A and 1B.

For the NFT sessions the EPI volumes were exported using Ice and Gadgetron[32]. In-house scripts created using Gadgetron and MATLAB (Mathworks) were used to reconstruct the 3D EPI data using SENSE[33] such that they could be read in near real-time by TBV to produce the target ROI time-series. There was a small delay at the start of each run to enable MATLAB to start, but after about 15s both the MRI scanner and the Gadgetron pipeline were fully in-synch with approximately 1s latency. To enable for both systems to synchronize we introduced a delay of 18 volumes at the start of each run. During this time participants viewed a white cross on a black background followed by a count-down (from 10 to 1) until the NFT paradigm started. In-house MATLAB scripts were used to process the ROI time-series and record participants' responses, breathing and heart rate. For the NFT runs, the ROI signal was regressed against head motion traces and physiological noise from respiration[34] and cardiac rhythm using RETROICOR[35]. The "cleaned" signal was then processed by in-house MATLAB scripts using Cogent toolbox (http://www.vislab.ucl.ac.uk/cogent_2000.php) to calculate and present the feedback to the participant. The computer setup in the scanner is shown in the Supplementary Figure 2B. Supplementary Figure 3 and Supplementary Figure 4 show the evoked response patterns and correlation coefficients respectively from the real-time processing pipeline. No statistical analyses were performed on these data, but are presented here for completeness. All analyses mentioned in the main document were performed on the data after extensive quality control and pre-processing (see section bellow on offline pre-processing).

Participants completed 4 NFT sessions across multiple sessions and each session included 4 NFT runs. Two participants completed 3 runs instead of 4 during one of the NFT sessions, because of fatigue. The activity NFT runs consisted of 6 rest blocks (30s), 5 response blocks (18s) and 5 upregulation blocks (30s). The rest and response blocks were identical to those of the transfer runs described above. During the upregulation blocks feedback was presented continuously in the form of a red bar. In the treatment group the height of the red bar represented the percent signal change at a given point during the upregulation blocks vs the mean activation during the preceding rest block. In the sham control group the height of the red bar was calculated using data from a yoked participant in the treatment group. A black line was set at 3/5 of the total bar height and acted as an additional reminder to the participants that they needed to increase the height of the red bar. Once the upregulation blocks started, there was an average delay of 2s until the red bar appeared on the screen and then it was updated roughly every 1.2s. The initial delay allowed for the ROI time-series

to be processed and the GLM model estimation to be initiated. The design of the task is shown in Figure 4C.

The connectivity NFT runs consisted of 5 rest blocks (45s), 5 upregulations blocks (30s) and 5 feedback blocks (3s). Feedback was presented intermittently at the end of the upregulation blocks in the form of a red bar. Similar to the activity NFT runs a black line was set at 3/5 of the total bar height as an additional reminder to the participants to upregulate. In the treatment group the height of the red bar was calculated using the Pearson's correlation coefficient between the SMA and left striatum ROI time-series during the upregulation blocks only[15]. In the sham control group the feedback was calculated using data from a yoked participant in the treatment group. The design of the task is shown in Figure 4D.

Similar to our previous study, we used shaping in order to facilitate learning and motivation[11,36,37], whereby the difficulty in increasing the height of the feedback bar was adjusted using the participants' performance in the preceding block. The details of the shaping approach are described fully in our previous study[11].

Prior to scanning, all participants were instructed to refrain from making any overt movements and only use mental strategies, such as motor imagery, in order to increase the levels of the NFT target (represented by the height of the red bar). Compliance was monitored using pneumatic tubes similar to our previous study[11]. More details are provided in the supplementary methods.

MRI Acquisition and Processing

Scanning Parameters

All scanning was performed on a Siemens TIM Trio 3T scanner using a standard 32-channel head coil. For the fMRI tasks we used a whole-brain multi-shot 3D echo-planar imaging (EPI) sequence[38] with TR = 1.2 s, TE = 30ms, excitation flip angle = 15°, bandwidth = 2604 Hz/Px. There were 60 slices per slab, acquired with sagittal orientation and anterior to posterior phase encoding. Image in-plane resolution was 64x64 and voxel size = 3x3x3 mm^3. To allow fast whole-brain coverage we used GRAPPA parallel imaging in phase encoding and partition encoding direction with 2x3 acceleration.

For the quantitative Multi-Parameter Maps[28,39,40] we acquired three spoiled multi-echo 3D fast low angle shot (FLASH) whole-brain volumes: (a) proton density (PD) weighted images with flip angle = 6°, TR = 25ms and with eight echoes TE=2.34ms, 4.64ms, 6.94ms, 9.24ms, 11.54ms, 13.84ms, 16.14ms and 18.44ms, (b) T1-weighted images with flip angle = 21°, TR = 25ms and with eight

TE=2.34ms, 4.64ms, 6.94ms, 9.24ms, 11.54ms, 13.84ms, 16.14ms and 18.44ms, and (c) magnetisation transfer (MT) weighted images with flip-angle = 6°, TR = 25ms and with six TE=2.34ms, 4.64ms, 6.94ms, 9.24ms, 11.54ms and 13.84ms. To achieve magnetisation transfer weighting an off-resonance RF pulse was applied before non-selective excitation. All volumes had 0.8mm isotropic voxel resolution, a field of view (FoV) of 256x224 mm^2 and 224 slices. To shorten acquisition time we used GRAPPA parallel imaging in phase encoding (anterior-posterior) and partition encoding (right-left) direction with 2x2 acceleration. Each scan lasted 7 mins. Prior to the acquisition of the MPMs, calibration data (B1 and B0 maps) were acquired to correct for inhomogeneities in the RF transmit field. MPMs were acquired only during the baseline and three follow-up sessions.

Offline fMRI Analysis

Statistical Parametric Mapping SPM12 (Wellcome Trust Centre for Neuroimaging, London) was used for offline pre-processing of the fMRI data. The first 3 volumes were removed from all fMRI time series apart from the NFT runs, where we removed the first 18 volumes. The images were then corrected for head-motion with rigid-body realignment using a 2-step approach and smoothed using an isotropic 8mm FWHM Gaussian smoothing kernel. Six motion parameters were generated for each fMRI run. In participants who used the PMCS the motion parameters represented residual and not actual head movement, i.e. they reflected motion that could not be fully corrected by the PMCS and remained in the time-series. For this reason, we did not use the motion parameters in order to identify and exclude bad scans, e.g. by examining scan-to-scan motion. Instead we used the DVARS[41] approach, the root mean square of the signal difference between consecutive scans, with the tool developed by Afyouni and Nichols[42], which provides a more standardized version of DVARS and computes p-values for a null hypothesis of homogeneity. Volumes were identified as bad if the change in DVARS was greater than 20% and were added in separate regressors to the first-level statistical models. Blocks for which more than half of the volumes were de-weighted were excluded from the condition of interest regressor (e.g. upregulation) and modelled as separate regressors. Data were also inspected for the presence of overt hand movements during upregulation or rest blocks, if a response was detected, the block was excluded from the condition of interest regressor and added to a separate task regressor. Runs with less than 2 blocks remaining in the upregulation condition were excluded from the analyses completely. Supplementary Table 5 shows the number of sessions that were included in the analyses across the groups for the 32 participants. In the case of

the participant who was excluded from the analyses and replaced, all the runs were contaminated by large head movements across the time-series and had to be rejected.

First-level, within-subject models included the condition of interest regressors. We used 2 regressors modelling the upregulation and response blocks for the baseline, NFT and transfer runs, and 1 regressor modelling the fist clenching blocks for the localiser runs. The baseline condition was modelled implicitly. In addition, models included 6 head motion parameter regressors produced by SPM and extracted from the PMCS (when used) with their temporal derivatives and the quadratic expansions of the parameters and their derivatives[43,44], spike regressors[45], as well as 13 physiological noise regressors modelling the heart rate using RETROICOR and respiration[34,35,46]. Temporal autocorrelation was modelled using SPM's first-level autoregressive process (AR(1)) and a high-pass filter with 128s cutoff.

For the activity NFT group, contrast values for upregulation vs baseline were extracted for the target ROI for each session and the highest 10% of t-values[47] were used to calculate the average ROI value. For the connectivity NFT group, the time-series for the target ROIs (SMA and striatum) was extracted using a 6mm sphere centred on the peak for upregulation vs baseline across all runs. The Pearson's correlation coefficient of the time-series between the two ROIs within the upregulation periods was then calculated and transformed into Fisher z-scores for analyses.

For the between-group comparisons of the training effect, the ROI estimates and transformed correlation coefficients were used as outcomes in linear repeated-measure models with group (treatment vs control) and session as fixed effects. Session was modelled as a repeated factor within subjects. Intersession covariance was modelled using heterogeneous compound symmetry (CSH), as this gave a reasonable approximation of the observed within-subject covariance while using minimal degrees of freedom. For the transfer effects, we tested change from baseline for each follow-up session separately and ran 1-way ANCOVAs with group as a fixed effect, adjusting for baseline levels of the NFT target to increase model sensitivity[18]. We used SAS 9.4 to estimate the repeated-measure linear models and ANCOVAs. For the repeated measure models using CSH, effect size was calculated using the effect t-statistic as the numerator and the square root of the sample size per group as the denominator. For the ANCOVAs, effect size was calculated using the contrast estimate as the numerator and the square root of the mean square error as the denominator.

Cognitive and Psychomotor Assessments

To assess change in cognitive and psychomotor function following neurofeedback training we calculated a composite score using the same measures and procedure as in our previous study[11]. This was a-priori specified using a set of independently validated measures sensitive to HD progression[5,6,48,49]. As a summary, the cognitive measurements included were: number correct for Stroop Word Reading only, number correct for Symbol Digit Modalities Test (SDMT), annulus length for Indirect Circle Tracing (log transformed) and number correct for negative Emotion Recognition. The Q-Motor measurements included were: inter-tap interval (ITI) and standard deviation of inter-onset interval (log transformed; log SD IOI) during speeded tapping with the left (non-dominant) index finger, and standard deviation of mid-tap interval deviation from target rhythm (log transformed; log SD dMTI) for paced tapping with left index finger at 1.8Hz. The composite score at the baseline session correlated highly with the normalized CAG Age Product score[50] (CAP score; Spearman's $r = -0.7$, $p < 0.001$) and the MOCA (Spearman's $r = 0.6$, $p < 0.001$) after controlling for age (results were also significant without controlling for age, both $p < 0.001$). It was therefore a sensitive measure of the participant's disease stage and overall cognitive and capacity.

References

1. Sitaram, R. et al. *Nat. Rev. Neurosci.* **18**, 86–100 (2017).
2. Novak, M.J.U. et al. *Hum. Brain Mapp.* **36**, 1728–1740 (2015).
3. McColgan, P. et al. *Brain* **138**, 3327–3344 (2015).
4. Poudel, G.R. et al. *Neurobiol. Dis.* **65**, 180–187 (2014).
5. Tabrizi, S.J. et al. *Lancet Neurol.* **10**, 31–42 (2011).
6. Tabrizi, S.J. et al. *Lancet Neurol.* **8**, 791–801 (2009).
7. Bartzokis, G. et al. *Neurochem. Res.* **32**, 1655–1664 (2007).
8. Emmert, K. et al. *NeuroImage* **124**, **Part A**, 806–812 (2016).
9. Horovitz, S.G., Berman, B.D. & Hallett, M. *Eng. Med. Biol. Soc. EMBC 2010 Annu. Int. Conf. IEEE* 4270–4273 (2010).doi:10.1109/IEMBS.2010.5627170
10. Ruiz, S. et al. *Hum. Brain Mapp.* **34**, 200–212 (2013).
11. Papoutsi, M. et al. *Hum. Brain Mapp.* **39**, 1339–1353 (2018).
12. Subramanian, L. et al. *J. Neurosci.* **31**, 16309–16317 (2011).
13. Subramanian, L. et al. *Front. Behav. Neurosci.* 111 (2016).doi:10.3389/fnbeh.2016.00111
14. Klöppel, S. et al. *Brain* **132**, 1624–1632 (2009).
15. Megumi, F., Yamashita, A., Kawato, M. & Imamizu, H. *Front. Hum. Neurosci.* **9**, (2015).
16. Yamashita, A., Hayasaka, S., Kawato, M. & Imamizu, H. *Cereb. Cortex* **27**, 4960–4970 (2017).
17. Hellrung, L. et al. *NeuroImage* **166**, 198–208 (2018).
18. Dimitrov, D.M. & Rumrill, J. *Work* **20**, 159–165 (2003).
19. Young, K.D. et al. *Am. J. Psychiatry* **174**, 748–755 (2017).
20. Ramot, M. et al. *eLife* (2017).doi:10.7554/eLife.28974

21. Wolf, R.C., Vasic, N., Schönfeldt-Lecuona, C., Ecker, D. & Landwehrmeyer, G.B. *Hum. Brain Mapp.* **30**, 327–339 (2008).
22. Domínguez D, J.F. et al. *Cortex* **92**, 139–149 (2017).
23. Foroughi, C.K., Monfort, S.S., Paczynski, M., McKnight, P.E. & Greenwood, P.M. *Proc. Natl. Acad. Sci.* **113**, 7470–7474 (2016).
24. Mehler, D.M.A. et al. *Neuropsychopharmacology* 1 (2018).doi:10.1038/s41386-018-0126-5
25. Yoo, S., Lee, J., O'Leary, H., Panych, L.P. & Jolesz, F.A. *Int. J. Imaging Syst. Technol.* **18**, 69–78 (2008).
26. Huntington Study Group *Mov. Disord.* **11**, 136–142 (1996).
27. Nasreddine, Z.S. et al. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
28. Weiskopf, N. et al. *Front. Neurosci.* **7**, (2013).
29. Nicholson, A.A. et al. *Hum. Brain Mapp.* **38**, 541–560 (2017).
30. Paret, C. et al. *Front. Behav. Neurosci.* **8**, (2014).
31. Paret, C. et al. *NeuroImage* **125**, 182–188 (2016).
32. Hansen, M.S. & Sørensen, T.S. *Magn. Reson. Med.* **69**, 1768–1776 (2013).
33. Pruessmann, K.P., Weiger, M., Scheidegger, M.B. & Boesiger, P. *Magn. Reson. Med.* **42**, 952–962 (1999).
34. Birn, R.M., Smith, M.A., Jones, T.B. & Bandettini, P.A. *NeuroImage* **40**, 644–654 (2008).
35. Glover, G.H., Li, T.-Q. & Ress, D. *Magn. Reson. Med.* **44**, 162–167 (2000).
36. Linden, D.E.J. et al. *PLoS ONE* **7**, e38115 (2012).
37. Weiskopf, N. et al. *J. Physiol.-Paris* **98**, 357–373 (2004).
38. Lutti, A., Thomas, D.L., Hutton, C. & Weiskopf, N. *Magn. Reson. Med.* **69**, 1657–1664 (2013).
39. Draganski, B. et al. *NeuroImage* **55**, 1423–1434 (2011).
40. Callaghan, M.F. et al. *Neurobiol. Aging* **35**, 1862–1872 (2014).
41. Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L. & Petersen, S.E. *NeuroImage* **59**, 2142–2154 (2012).
42. Afyouni, S. & Nichols, T.E. *NeuroImage* **172**, 291–312 (2018).
43. Ciric, R. et al. *NeuroImage* **154**, 174–187 (2017).
44. Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S.J. & Turner, R. *Magn. Reson. Med.* **35**, 346–355 (1996).
45. Lemieux, L., Salek-Haddadi, A., Lund, T.E., Laufs, H. & Carmichael, D. *Magn. Reson. Imaging* **25**, 894–901 (2007).
46. Hutton, C. et al. *NeuroImage* **57**, 101–112 (2011).
47. Todd, N. et al. *Front. Neurosci.* **11**, (2017).
48. Tabrizi, S.J. et al. *Lancet Neurol.* **11**, 42–53 (2012).
49. Tabrizi, S.J. et al. *Lancet Neurol.* **12**, 637–649 (2013).
50. Ross, C.A. et al. *Nat. Rev. Neurol.* **10**, 204–216 (2014).

Acknowledgments

We are very grateful to the study participants and their families for their support. We would also like to thank the imaging support team at the Wellcome Centre for Human Neuroimaging and our funders. This study was funded by an MRC Developmental Pathway Funding Scheme grant (MR-L012936-1), a European Huntington's Disease Network (EHDN) seedfund grant, and the Wellcome Trust. ST is funded by a UCLH BRC grant, GR is funded by a Wellcome Trust Senior Research Fellowship, and NW is funded by WTCN Core funding.

Conflict of Interest

The Wellcome Centre for Human Neuroimaging and the Max Planck Institute for Human Cognitive and Brain Sciences have institutional research agreements with and receive support from Siemens Healthcare.

Data Availability

All data are available from the authors and cannot become publicly available due to lack of consent from the study participants.