UNIVERSITY OF LONDON THESIS

Degree PHD Year 2008 Name of Author HENLEY, SUSIE, MAYCHELINE, DEBORAH.

MRI and cognitive changes in Huntington's Disease

Thesis submitted for the degree of Doctor of Philosophy

Susie M. D. Henley

Institute of Neurology

University College London

2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I, Susie Henley, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

ABSTRACT

This thesis focused on cognitive and MRI measures of early change and progression in Huntington's disease (HD). HD is clinically heterogeneous and previous findings about the location of brain atrophy in the early stages, and its relation to cognition, are equivocal. A pilot study assessed the practicality of using serial volumetric MRI in early HD and the usefulness of cognitive tasks in longitudinal assessment. A larger study investigated cross-sectional and longitudinal aspects of these domains in premanifest and early HD.

Global and regional cerebral volumes were investigated using manual volumetry and voxel-based morphometry (VBM). The work describes the application of the Brain Boundary Shift Integral (BBSI) to measure of whole-brain atrophy rate. Preparatory technical work included choice of templates for optimum scan alignment and segmentation in VBM, and adjustment of BBSI parameters to obtain maximum agreement between it and manual measures. Cognitive ability was assessed using a wide-ranging battery of tests, some standard and some novel.

Atrophy in early HD was found to be more extensive than previously reported, involving widespread extra-striatal loss, and not all functional deficits could be attributed to striatal damage. Emotion recognition deficits were broad and associated with striatal and extra-striatal brain regions. Executive function and memory tasks demonstrated decline over one year. Whole-brain atrophy rates were increased in early HD and associated with decline in cognition and CAG repeat length.

This work elucidates the extent of, and associations of, atrophy and cognitive impairment in HD, and adds weight to the suggestion that variability in progression rates is partly explained by genetic factors. The suggestion that motor learning might be impaired even in very far-from-onset gene carriers, and that tasks other than those tapping executive function were sensitive to decline, should motivate further work aimed at detecting the very earliest signs of change in this disease.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF TABLES IN APPENDICES

LIST OF FIGURES IN APPENDICES

LIST OF ABBREVIATIONS

AIR	Automated Image Registration
ANART	American National Adult Reading Test
BBSI	Brain boundary shift integral
BDI	Beck Depression Inventory
BOLD	Blood oxygen level dependent
CAG	Cytosine adenine thymine, the codon that encodes the amino acid glutamine
CI	Confidence interval
CSF	Cerebrospinal fluid
CT	Computerised tomography
dof	Degrees of freedom
DWI	Diffusion weighted imaging
EHDN	European Huntington's Disease Network
FDR	False discovery rate
fMRI	Functional magnetic resonance imaging
FWE	Family-wise error
FWHM	Full width at half maximum
GM	Grey matter
GNT	Graded Naming Test
HD	Huntington's disease
HMGT	Homophone meaning generation test
HVLT	Hopkins Verbal Learning Test
IQ	Intelligence quotient
MIDAS	Medical Information Display and Analysis System
MMSE	Mini-Mental State Examination
MNI	Montreal Neurological Institute
MRI	Magnetic resonance imaging
NART	National Adult Reading Test
PBA-HD	Problem Behaviors Assessment for Huntington's Disease
PET	Positron emission tomography
PM	Premanifest

rCBF	Regional cerebral blood flow
Resel	Resolution element (a measure of the spatial resolution of an image in which voxels are correlated)
RMF	Recognition memory for faces (subtest of RMT)
RMT	Recognition Memory Test
RMW	Recognition memory for words (subtest of RMT)
ROI	Region of interest
SBA	Short Behavioural Assessment
SD	Standard deviation
SDMT	Symbol Digit Modalities Test
SPM	Statistical parametric map
SPMx	Statistical Parametric Mapping (version number): software used for voxel-based morphometric analyses
TFC	Total Functional Capacity
TIV	Total intracranial volume
TMT	Trail Making Test
UHDRS	Unified Huntington's Disease Rating Scale: the clinical assessment tool developed for use in HD
VBM	Voxel-based morphometry
VOSP	Visual Object and Space Perception Battery
WAIS	Wechsler Adult Intelligence Scale
WM	White matter

1 HUNTINGTON'S DISEASE

1.1 BACKGROUND AND HISTORY OF HUNTINGTON'S DISEASE

One of the earliest descriptions of "hereditary chorea" is ascribed to George Huntington in 1872 (Huntington 1872) based on his observations of a set of families on Long Island with abrupt, rapid involuntary movements of the body, or chorea. In these families he noted that the chorea was hereditary, that onset was "invariably" between the ages of 30 and 40, and that there was a tendency towards insanity or suicide as the disease progressed and the mind became more impaired. He spoke of the disease ". . . hardly ever manifesting itself until adult or middle life, and then coming on gradually but surely, increasing by degrees, and often occupying years in its development, until the hapless sufferer is but a quivering wreck of his former self." He wrote that he knew nothing of its pathology and ended the paper saying ". . . I have never known a recovery or even an amelioration of symptoms in this form of chorea; when once it begins it clings to the bitter end. No treatment seems to be of any avail, and indeed nowadays its end is so well-known to the sufferer and his friends, that medical advice is seldom sought." (Huntington 1872).

George Huntington was describing what is now known as Huntington's disease (HD), which, as he so astutely noted over 100 years ago, is an autosomal dominant neurodegenerative disorder usually of adult onset. Although for many years it was characterised primarily as a movement disorder, it is now well recognised that cognitive and psychiatric problems accompany and frequently precede motor signs. Early in the disease psychomotor and executive skills (such as planning or set switching) tend to be affected, and depression, apathy or

irritability are some of the more common behavioural disturbances (Craufurd and Snowden 2002; Thompson *et al.* 2002). It has a prevalence in the UK of ~1 per 10,000 individuals with many more at risk of the disease (Harper 2002). It progresses slowly and inexorably, with increasingly distressing and disabling symptoms associated with considerable care-giver and family burden, with morbidity and death occurring 15-20 years from onset, by which time the patient is often bed-bound and mute. Much more is now known about the pathology of the disease, and the gene that causes it, than in George Huntington's day. However, although a number of symptoms can be alleviated there is still no treatment known to stop or even slow the course of the disease.

1.1.1 The HD gene and pathology

In 1993 the genetic abnormality underlying HD was discovered to be a CAG expansion on chromosome 4 (Huntington's Disease Collaborative Research Group 1993). Alleles of up to 35 repeats are normal, whilst alleles with 40 or more repeats are fully penetrant and the carrier is likely to show signs of the disease within a normal lifespan. The intermediate repeat numbers (36-39) are not fully penetrant but there have been reports of 36 CAG repeats leading to the disease, and of people living into their 90s with 39 CAG repeats and no signs of HD (Rubinsztein *et al.* 1996). Higher CAG repeat lengths are associated with earlier onset although CAG repeat length only explains about 50-70% of the variance in age at onset and other genetic or environmental factors are likely to influence this (Duyao *et al.* 1993). The repeat is unstable and can change during transmission, with larger expansions particularly likely to occur in transmission from the male. Non-abnormal or intermediate repeat lengths can also expand

into the pathogenic range, leading to new mutations in previously unaffected families (Bates 2005).

The gene encodes the protein huntingtin which is expressed widely throughout the brain and central nervous system. The precise role of huntingtin is unclear but it appears to be multifunctional, and it is generally believed that the symptoms seen in HD are caused by toxicity of the mutant form of the protein (Wanker and Droge 2002) although there is some evidence that reduced wild-type huntingtin function might also contribute to the disease process (Cattaneo *et al.* 2005). The presence of mutant huntingtin leads to the selective neuronal dysfunction and death which underlie the clinical features of the disease.

The most striking atrophy in HD is seen in the striatum (caudate nucleus and putamen), in which atrophy progresses along a dorso-medial – ventro-lateral gradient (Vonsattel *et al.* 1985). At post-mortem extensive and relatively uniform atrophy is also seen throughout the cortex, with probably slightly more loss of white matter relative to grey matter (de la Monte *et al.* 1988; Mann *et al.* 1993). It is not yet clear whether cortical atrophy is secondary to striatal atrophy or an independent process, nor why some neurons are selectively vulnerable in HD, although this may be in part due to mutation length gains within some cell types and regions (Shelbourne *et al.* 2007). Brain volume at post-mortem is almost 20% less than in controls, and cortical volume has been shown to correlate with striatal volume, suggesting that linked degenerative processes might underlie the loss in both regions (Halliday *et al.* 1998). Further structural and functional abnormalities that can be seen *in vivo* are discussed in section 1.2.

1.1.2 Testing, diagnosis and treatment

The discovery of the causative gene made it possible for the presence or absence of the abnormal allele to be confirmed by genetic testing. Patients showing signs of HD can be offered a confirmatory test, and adults at risk of HD by virtue of having an affected parent can be offered predictive testing, although this needs to be approached extremely carefully and currently only between 10 and 25% of those eligible to have the test do so, perhaps because of the knowledge that there is, as yet, no cure (see Tibben 2002).

Clinical "onset" of HD is not diagnosed until the clinician is confident that the patient is showing unequivocal motor signs. However, as mentioned above, it is not uncommon for subtle motor, cognitive or behavioural abnormalities to be present many years prior to this (e.g., Lawrence *et al.* 1998a; Williams *et al.* 2007). Motor symptoms such as abnormal eye movements may also be present in premanifest gene carriers (Golding *et al.* 2006). Depression and other affective disorders can precede motor symptoms by many years, and this is not merely a response to the situation in which carriers of the abnormal gene find themselves because it is the case even in those who do not know that they are at risk of the disease (Craufurd and Snowden 2002; Julien *et al.* 2007). Cognitive symptoms are described in more detail in section 1.3.

As yet there are no pharmaceutical interventions known to delay or prevent onset of HD although symptoms such as chorea and depression can be alleviated. A number of groups are testing potential neuroprotective compounds, some of which have been shown to be safe and well-tolerated, although larger studies are needed to see if they can demonstrate any effect on function. Other strategies

such as striatal grafts and gene therapy are also being considered although currently these approaches are somewhat held back by both ethical and technical issues and so more research needs to be done if either approach is to become useful for treating HD (Handley *et al.* 2006).

1.1.3 Current clinical measures in HD

Therapeutic trials in HD investigating disease-modifying effects are challenging for a number of reasons in addition to those mentioned in the previous section, predominantly the difficultly in measuring and quantifying disease-related decline over time, and by extension, the difficulty in determining whether the rate of that decline has been slowed by an intervention. This is primarily because of the heterogeneity of the disease, and the inadequacy of the current clinical measurement scales, particularly in the very earliest stages of the disease.

In HD, disease progression is relatively slow, and the disease is highly clinically heterogeneous, so patients with HD present with variable clinical phenotypes (differing burden of motor, cognitive and psychiatric symptoms). There is also a suggestion that higher CAG repeat length is associated with faster clinical progression, and that this interacts with time since diagnosis (Rosenblatt *et al.* 2006). This inherent variability and non-linearity makes it hard to detect small differences between groups, or change over time.

The current clinical rating scales are the Unified Huntington's Disease Rating Scale (UHDRS) and Total Functional Capacity (TFC) rating scales (Shoulson and Fahn 1979; Huntington Study Group 1996) and they are widely used, but have high measurement variability, particularly in the behavioural/psychiatric assessment (Craufurd and Snowden 2002). The motor subscale of the UHDRS

shows only fair agreement between clinicians (Hogarth *et al.* 2005) and the authors recommend that investigators be selected and trained carefully if they are to perform such assessments in therapeutic trials. Because of this variability these scales can require long assessment intervals and large subject numbers in order to have the statistical power to detect effects on function (Kieburtz and Shoulson 2002) and importantly the scales have limited ability to distinguish effects on disease progression from symptomatic benefit. They are also subject to floor and ceiling effects, with premanifest and even early disease subjects unlikely to score below ceiling for the TFC and independence scales, and those in the final stages of the disease likely to be immobile and totally dependent and thus scoring at floor level.

As mentioned above, disease onset is defined as the time at which a clinician confirms the unequivocal presence of motor signs, but subtle motor signs, as well as cognitive and behavioural changes, can be present prior to this in the so-called "premanifest" population. Therapies that delay onset or slow functional decline are likely to be targeted at premanifest or very early HD patients. However, on standard tests most premanifest subjects perform in the clinically normal range (Paulsen *et al.* 2006a; Solomon *et al.* 2007) and, as discussed, current clinical measures are insensitive in both premanifest and very early HD populations. Inevitably clinical scales measuring "symptoms" will be completely inadequate for measuring progression in subjects who are asymptomatic.

Therefore it is clear that more reliable, validated measures are needed, both to define and predict disease onset, and to quantify and track change over time accurately. Such measures would give a clearer picture of the disease,

particularly the functional and anatomical changes in the premanifest and early stages, and could potentially be used as markers of disease progression or surrogate endpoints against which to test the efficacy of therapies. The following sections summarise findings in neuroimaging, and cognition, both of which have the potential to provide more information about these aspects of HD.

1.2 NEUROIMAGING IN HD

Brain imaging can provide a non-invasive assessment of brain structure, function, metabolites or perfusion, and through the use of radio-labelled ligands allows estimates of specific receptor (e.g., Dopa) or molecular (e.g., amyloid) densities. Imaging is increasingly being used in neurological disorders both for diagnosis and for locating regions of dysfunction or atrophy and measuring progression. It can be repeated over time to give an idea of where change is occurring, and at what rate. It is also useful to visualise early brain abnormalities in the period prior to clinical diagnosis, i.e., at a point where clinical, behavioural or functional abnormalities are not noticeable or are too subtle for detection with standard clinical scales. In HD, because a genetic marker exists, imaging work has focused on predicting clinical onset from structural or functional abnormalities, and measuring rate of decline.

1.2.1 Imaging modalities

Structural MRI uses the signal emitted by protons in different brain tissues, when aligned in a magnetic field and exposed to a radio-frequency pulse, to reconstruct an image of those tissues. Parameters can be altered to adjust the relative signal intensities from different tissue types, in order to highlight particular types of pathology, or to maximise the contrast between tissues. Structural MRI is widely

used in studies of neurodegeneration because it is possible to visualise patterns of atrophy, and to quantify the volume of different structures from the MR image.

Quantification often takes the form of manual or semi-automated outlining of a structure of interest on an image in order to estimate its volume (a region of interest, or ROI approach). A different approach which has been used more frequently recently is to examine whole tissue types and look for regions where statistically significant differences in the image intensities exist between groups (cortical thickness measurement, and voxel-based morphometric methods are examples of this). The latter approaches do not require *a priori* hypotheses about the location of atrophy and are thus considered "unbiased" ways of examining morphometric changes across the whole brain (see sections 1.2.2 and 2.5.5). Manual ROI approaches are usually preferred for generating reliable estimates of the volume of specific anatomical structures.

Other imaging techniques allow measurement of a number of different parameters including: regional cerebral blood flow (rCBF); brain metabolic activity; brain metabolites; brain receptor binding or water diffusivity as indices of brain function, either in a resting state or during a behavioural task. They have the potential to detect functional change before clinical or volumetric changes are apparent and thus could be particularly useful as markers of "premanifest" disease onset, and as measures of potential neuroprotective therapies. Functional MRI (fMRI) measures the blood oxygen level dependent (BOLD) haemodynamic response in the brain, and thus is an indirect measure of neuronal activity. Position emission tomography (PET) uses radiotracers to measure rCBF, metabolism or brain receptor binding. Diffusion weighted

imaging (DWI) measures the diffusivity of water in the brain which can be used as an index of structural integrity.

The imaging focus of this thesis is on structural MRI, and in order to set this in context the following sections give an overview of relevant structural and functional imaging findings in HD. These sections are not intended to be comprehensive.

1.2.2 Structural MRI

1.2.2.1 Cross-sectional studies

As mentioned in section 1.1.1 the most striking pathological changes in HD occur in the striatum (Halliday *et al.* 1998; Gutenkst *et al.* 2002) and a number of cross-sectional structural MRI studies in HD have demonstrated atrophy of the putamen and caudate nucleus *in vivo* (Harris *et al.* 1992; Harris *et al.* 1996; Rosas *et al.* 2001). In these studies the putamen tended to show stronger negative associations with motor and chorea scores (Harris *et al.* 1992) although volumes of both structures were associated with cognitive performance, (Harris *et al.* 1996) and total striatal loss was predicted by CAG repeat length (Rosas *et al.* 2001).

Loss of striatal volume has also been reported in premanifest HD gene carriers, demonstrating that striatal neuronal loss occurs even before onset of clinical symptoms and that caudate and putamen volumes decrease with proximity to predicted motor onset (Aylward *et al.* 1994; Aylward *et al.* 1996; Harris *et al.* 1999). In one study caudate volume alone predicted with 100% accuracy subjects who would be diagnosed within two years (Aylward *et al.* 2004). Reduced putamen volume has been found in subjects as much as 23 years from

predicted motor onset, although caudates in this group were not significantly smaller than in controls (Paulsen *et al.* 2006b).

These volumetric imaging studies in HD were conducted to analyse volume changes in predefined pathologically-implicated cerebral regions of interest, solely in subcortical structures. However there is increasing post-mortem evidence that atrophy is more widespread, and that subcortical atrophy might be associated with cortical changes (de la Monte *et al.* 1988; Halliday *et al.* 1998). Similar changes have also been reported *in vivo*. Generalised white matter loss has been reported in early HD, in the absence of total brain volume or grey matter loss, with white matter volume associated with psychomotor and executive task performance (Beglinger *et al.* 2005). Disproportionate frontal lobe white matter loss has also been demonstrated, with no differences in overall brain volume until patients were moderately affected (Aylward *et al.* 1998). However others have reported reduced total brain volume in early HD (Rosas *et al.* 2003; Kassubek *et al.* 2004a). A detailed study found grey matter loss in the frontal and temporal lobes, and white matter loss in deep, frontal and occipital regions, in mild to moderate HD (Fennema-Notestine *et al.* 2004). These authors also found significant associations between caudate volume and cortical grey and white matter volume, suggesting that a similar or parallel neurodegenerative mechanism might underlie both striatal and cortical atrophy.

These findings, which are based on volumetric measurements of structures of interest, are reinforced by the finding of reduced cortical thickness in early to moderate HD, with a suggestion that the sensorimotor areas are most severely

affected and that atrophy progresses from posterior to anterior regions with longer disease duration (Rosas *et al.* 2002; Rosas *et al.* 2008).

Abnormal brain volume has also been found in premanifest subjects. One study found evidence of reduced white matter and normal grey matter volume in a premanifest group predicted to be ~11 years from onset. In contrast a premanifest group ~23 years from onset had increased grey matter, and decreased white matter, in the absence of any increase in ventricular size or cerebrospinal fluid (CSF) volume, which led the authors to postulate that this might represent a developmental abnormality rather than a very early neurodegenerative process. Total brain volume was not significantly reduced (Paulsen *et al.* 2006b) although further work on the same cohort demonstrated morphological grey matter abnormalities, including increased gyral surface area, thicker gyri and thinner sulci (Nopoulos *et al.* 2007). Decreased white matter has been reported to correlate with years to onset in premanifest subjects, and in this case grey matter was also reduced, although not significantly (Ciarmiello *et al.* 2006). However grey matter thinning in regions across the cortex has also been reported in this population (Rosas *et al.* 2005). In this study grey matter thickness in some regions correlated with striatal volumes, and thinning was also associated with worse performance on psychomotor and executive task performance.

The technique of voxel-based morphometry (VBM) (Ashburner and Friston 2000) is increasingly used in the study of neurodegeneration, both to determine patterns of atrophy and to analyse associations between atrophy and behaviour. As mentioned above this technique is an unbiased, automated method for whole-

brain analysis (see section 2.5.5 for more details). Its advantage over ROI techniques is that it requires no *a priori* hypotheses about particular regions of interest. VBM therefore has the potential to detect atrophy in areas that can be missed by ROI studies.

VBM findings in HD are mixed; most studies find evidence of bilateral caudate and putamen atrophy in both early and premanifest HD (Thieben *et al.* 2002; Peinemann *et al.* 2005). In contrast there is less consensus about extrastriatal atrophy; for example insular atrophy has been shown in premanifest (Thieben *et al.* 2002), but not in some of the early HD studies (Kassubek *et al.* 2004c). There are similarly variable results on the involvement of the thalamus, amygdala, hypothalamus and frontal areas (Kassubek *et al.* 2004c; Kassubek *et al.* 2005; Peinemann *et al.* 2005; Douaud *et al.* 2006; Mühlau *et al.* 2007b).

In summary, structural MRI studies in early HD tend to support the post-mortem findings of widespread generalised atrophy throughout the cortex, with disproportionate volume loss in the striatum. Where both grey matter and white matter are measured, there is a suggestion that white matter loss is relatively greater or seen earlier in the disease. Premanifest subjects have clear striatal atrophy and there is evidence of white matter loss but the involvement of grey matter is less clear. There is considerable variation between studies both with regard to the extent of extra-striatal atrophy and the point in the disease at which loss in different regions starts to become significant. In many cases these differences might be due to methodological differences, or lack of statistical power. Alternatively they may reflect true heterogeneity in the populations sampled. Overall this section demonstrates that although it is clear that

widespread structural losses occur early in the disease, the pattern, time course, severity, significance and evolution are all still unclear with some conflicting claims in the literature.

1.2.2.2 Longitudinal studies

The majority of longitudinal structural MRI studies in HD have focused on the striatum. Early evidence was that in close-to-onset and early HD striatal atrophy rates were fairly uniform across stages, but that higher CAG repeat lengths (associated with earlier age at motor onset) were associated with greater atrophy rates (Aylward *et al.* 1997; Aylward *et al.* 2000). It was later reported that the putamen and caudate began to atrophy at rates significantly greater than zero up to 11 years before disease onset with evidence of some non-significant decline between 20-10 years prior to onset, i.e. that atrophy rates were non-linear, at least in the premanifest stage (Aylward *et al.* 2004). These conclusions were based on "snapshots" of subjects at different years to onset, and atrophy rates were not shown to be greater than those seen in normal ageing, so longer prospective studies are needed in order to clarify the exact way in which atrophy rates change over time. Rate of caudate atrophy has been suggested as a surrogate marker of disease progression (Aylward *et al.* 2003; Aylward 2007).

There have been few other longitudinal structural MRI studies in HD. One study looked at total grey, white matter and CSF and found that these volumes did not change significantly over 18 months in early HD although white matter volume decreased in premanifest subjects over 16 months (Ciarmiello *et al.* 2006). VBM approaches have yielded slightly different results: a two-year study of premanifest subjects found evidence of subcortical grey matter atrophy, but no

changes in white matter (Kipps *et al.* 2005). In contrast a one-year study of early HD subjects found change in these regions as well as the insula, cerebellum and some cortical regions, although whether these changes are specific to HD is unclear as the age ranged from four to 73 but results were not adjusted for either age or head-size (Ruocco *et al.* 2007).

Overall recent evidence suggests that it may be possible to detect volume loss over time in regions outside the striatum but little work has focussed on quantifying this in more detail.

1.2.2.3 Effect of CAG repeat length on brain volume

CAG repeat length is known to account for between 50 – 70% of the variance in age at onset (Duyao *et al.* 1993; Langbehn *et al.* 2004) but there is debate as to what extent it contributes to other aspects of HD, including brain atrophy.

Some post-mortem studies have found no effect of CAG repeat length on cortical or sub-cortical cross-sectional areas (Sieradzan *et al.* 1997) whereas others find that CAG repeat length either predicts the ratio of cell loss to age at death (i.e. rate of cell loss) in the striatum (Furtado *et al.* 1996; Penney, Jr. *et al.* 1997), or atrophy rates in the cortex but not subcortical structures (Halliday *et al.* 1998). In contrast to the latter finding, *in vivo* MRI studies have shown that CAG repeat length is correlated with striatal but not cortical volumes in manifest HD patients (Ruocco *et al.* 2006) (unadjusted for age) and predicts age-adjusted striatal volume loss (Rosas *et al.* 2001) as well as rate of striatal loss (Aylward *et al.* 1997).

VBM approaches to this question have also yielded mixed results: an effect of CAG repeat length was found only in the right inferior frontal sulcus in premanifest subjects (Thieben *et al.* 2002), and of two studies which investigated the subcortical grey nuclei on a voxel-by-voxel basis in early HD, one found no relationship (Douaud *et al.* 2006) whilst the other found a negative association between CAG repeat length and caudate volume (Peinemann *et al.* 2005). Another VBM study examined this indirectly in early HD patients, comparing a "low" CAG group with controls, and a "high" CAG group with controls, and showed that the high CAG group had more extensive atrophy of the striatum but both groups showed similar amounts of extrastriatal atrophy (Kassubek *et al.* 2004c). However the two groups were not compared directly. In early HD, higher CAG repeat length has been associated with smaller striatal and sub-postcentral gyri white matter volume (Jech *et al.* 2007) and with increased extent and rate of atrophy in extra-striatal, but not striatal regions (although this was not corrected for age) (Ruocco *et al.* 2007).

Overall, most studies suggest that there is an influence of CAG repeat length on the location and amount of brain atrophy in HD but there are discrepancies as to exactly which regions are affected.

1.2.3 Functional imaging

Both hypoperfusion and (possibly compensatory) hyperperfusion have been demonstrated in early HD subjects. Reduced BOLD responses in frontal and occipital regions were found using an implicit learning task (Kim *et al.* 2004). Increased activation in the anterior cingulate, insula and frontal and parietal regions was shown in an executive function task (Georgiou-Karistianis *et al.*

2007) and this group has also reported relatively reduced functional connectivity in prefrontal regions using the same task (Thiruvady *et al.* 2006).

Reduced activity has also been shown in premanifest gene carriers, sometimes in the absence of measurable clinical or cognitive deficits (Reading *et al.* 2004; Hennenlotter *et al.* 2004). Paulsen *et al.* (2004) found that close-to-onset (<12 years) gene carriers had smaller caudate nuclei, worse time discrimination performance and reduced activation in the caudate, thalamus and putamen relative to controls. They were also perhaps the first to show reduced activation in the same areas in a group of gene carriers whose estimated time to onset was >12 years, in the absence of any volumetric or behavioural differences. Similar effects have been shown with a working memory task, with hypometabolism of the left dorsolateral prefrontal cortex related to working memory load in relatively far-from-onset premanifest subjects, again in the absence of behavioural or volumetric changes (Wolf *et al.* 2007), and a time reproduction task (Zimbelman *et al.* 2007).

PET studies have shown decreased striatal glucose metabolism and dopamine receptor binding in premanifest and early HD, with reduction in both measures associated with proximity to onset in premanifest subjects (Lawrence *et al.* 1998c; Ciarmiello *et al.* 2006; Feigin *et al.* 2007). Dopamine receptor binding may be a more sensitive marker of dysfunction in HD than glucose metabolism (van Oostrom *et al.* 2005), and increased rates of decline in striatal receptor binding have been shown in both premanifest and early HD subjects (Andrews *et al.* 1999; Pavese *et al.* 2003) with relatively linear change over three or four years (Pavese *et al.* 2003; Feigin *et al.* 2007). In early HD progressive reduction

of glucose metabolism and dopamine receptor binding has also been reported in the amygdala, frontal and temporal cortices (Pavese *et al.* 2003), and white matter (Ciarmiello *et al.* 2006), whilst hypermetabolism of the thalamus has been suggested as a possible compensatory mechanism in far-from-onset premanifest subjects (Feigin *et al.* 2007). PET has also been used as a marker of striatal graft treatment (Gaura *et al.* 2004; Furtado *et al.* 2005).

DWI studies have found compromised tissue integrity in early and premanifest HD in subcortical grey matter as well as white matter (Mascalchi *et al.* 2004; Reading *et al.* 2005), with increases in white matter diffusivity associated with decreased cognitive performance (Rosas *et al.* 2006). Compared with control subjects, premanifest subjects have a smaller percentage of white matter tracts from frontal areas to the caudate and this correlates with their performance on voluntary-guided saccades (Klöppel *et al.* 2008).

Functional imaging techniques have demonstrated widespread abnormalities in early and premanifest HD, and it is of particular note that compensatory mechanisms have been demonstrated, and that functional decline can be shown prior to behavioural or volumetric abnormalities. Functional imaging therefore shows promise as a tool for detecting the earliest premanifest change in HD and could be used longitudinally to measure rates of change.

1.2.4 Summary

As structural MRI techniques have moved from hypothesis-driven ROI approaches to more exploratory analyses of the whole brain, the concept of HD as a primarily striatal disorder has changed. Although atrophy of the striatum appears most severe and is apparent very early on, widespread cortical grey and

white matter loss, in both premanifest and early HD, has also been shown. The exact extent and time course of volume changes, and how they relate to CAG repeat length is unclear. Functional techniques have shown putative markers of dysfunction in HD, in some cases before volumetric or behavioural change was detectable. In one study functional changes showed little overlap with volumetric changes, suggesting that they were reflecting reduced striatal input to cortical regions, rather than structural changes (Gavazzi *et al.* 2007).

All these measures have been suggested as possible markers of disease progression, given that they are sensitive to abnormalities in HD and not subject to floor, ceiling, or practice effects as so many clinical scales are. Analyses can be shown to be highly reproducible and can be performed blinded to subject characteristics.

Structural imaging with MRI is non-invasive, widely available and relatively fast. Volumes and rates of change can be easily and reliably quantified. It is apparent from previous work that structural change in HD is extensive and progressive, but more needs to be known about the patterns of atrophy and how they develop in the early stages. In addition, it is clear that atrophy in HD is not restricted to the striatum, so it may be that measures that encompass the whole brain are also sensitive markers, at least at certain points in the disease. The majority of studies have been cross-sectional and so future work would benefit from longitudinal analysis of changes across the brain.

1.3 COGNITION IN HD

Many studies have examined cognitive deficits in HD with a view to clarifying the nature of the disorder and the problems that people with HD face. In relation

to the current need for reliable markers of disease progression, recent large-scale studies are starting to focus not only on which skills are impaired in HD, but on which tests are the most sensitive to deficits and to decline over time. The following review therefore focuses on the sensitivity and usefulness of tests as much as on what they can tell us about function in HD.

Early HD patients tend to be impaired at psychomotor tasks, executive function, with some perceptual problems and mild memory problems, whilst language comprehension, semantic memory and skill learning are considered relatively unaffected until later in the disease process (Bachoud-Lévi *et al.* 2001; Craufurd and Snowden 2002). Performance also tends to be worse in patients with longer CAG repeat lengths or disease duration (Jason *et al.* 1997; Rosenblatt *et al.* 2006). In premanifest gene carriers deficits are less obvious, and although subtle changes can be detected many gene carriers perform within normal limits (Paulsen *et al.* 2006a).

1.3.1 Cross-sectional studies

1.3.1.1 IQ

IQ as measured by the Wechsler Adult Intelligence Scale (WAIS) (Wechsler 1981) tends to be below the normal range in early manifest HD, with impairment on most subtests where they are reported (e.g., Josiassen *et al.* 1983; Giordani *et al.* 1995; Snowden *et al.* 2001) (but see also de Boo *et al.* 1997). Findings are less clear cut in premanifest HD, with reports of both normal function (de Boo *et al.* 1997; Lemiere *et al.* 2002) and an overall impairment (Giordani *et al.* 1995; Witjes-Ane *et al.* 2003). Findings on individual subtests are also mixed with some premanifest cohorts impaired only on picture completion (Giordani *et al.*

1995) and others showing problems with arithmetic and digit symbol (Kirkwood *et al.* 2000), as well as a version of picture arrangement (Snowden *et al.* 2002).

A French-language estimate of premorbid IQ found early HD patients to be in the normal range (Bachoud-Lévi *et al.* 2001), and this has also been demonstrated using the National Adult Reading Test (NART) (Snowden *et al.* 2002). However a large study of premanifest subjects found that errors on the ANART (the American-language version of the NART) increased slightly with proximity to estimated motor onset, which suggests that these tests may not be entirely insensitive to disease effects (Solomon *et al.* 2007).

1.3.1.2 Executive function

Patients with HD often demonstrate broad impairments in executive function (planning, shifting attention, decision-making, monitoring performance) and psychomotor tasks. Performance on the Stroop task is generally impaired in early HD (Giordani *et al.* 1995; de Boo *et al.* 1997; Snowden *et al.* 2001) although in one study the deficit was only shown on the more automated parts of the task (colour- and word-reading) with sparing of the interference part (Bachoud-Lévi *et al.* 2001). Similarly early HD patients often perform poorly on phonemic fluency and the Symbol Digit Modalities Test (SDMT) (Snowden *et al.* 2001; Lemiere *et al.* 2002), which together with the Stroop test comprise the cognitive parts of the UHDRS. However, not all studies find that phonemic fluency is impaired (de Boo *et al.* 1997; Bachoud-Lévi *et al.* 2001) and mixed findings are also reported for category fluency (Bachoud-Lévi *et al.* 2001; Lemiere *et al.* 2002). It has recently been suggested that impairments in fluency tasks in HD are commensurate with subjects' verbal ability and motor

impairments, and thus might reflect dysfunction in these areas rather than a disproportionate impairment in executive function *per se* (Henry *et al.* 2005); this might explain why not all studies have found these tasks to be impaired in the early stages.

Impairment on the SDMT can also be detected in premanifest subjects (Paulsen *et al.* 2001; Witjes-Ane *et al.* 2003) but the other tasks discussed above tend to be performed within the normal range by this group (e.g., Giordani *et al.* 1995; Lemiere *et al.* 2002). However some deficits have been shown in phonemic fluency (de Boo *et al.* 1997), category fluency (Lawrence *et al.* 1998a; Verny *et al.* 2007) and the Stroop test (Paulsen *et al.* 2001; Snowden *et al.* 2002; Verny *et al.* 2007) although in the latter two studies subjects were impaired only at the colour-reading subtest, not word-reading or interference. Deficits in attentional set-shifting have also been shown, in the absence of concurrent problems with planning and working memory (Lawrence *et al.* 1998a).

1.3.1.3 Psychomotor

Craufurd and Snowden (2002) point out that psychomotor slowing distinguishes sub-cortical from cortical dementias and is prominent in HD. It is often demonstrated on tasks such as Stroop and SDMT (discussed above) as well as Trail Making tests, performance on which is usually impaired in manifest patients (e.g., Bachoud-Lévi *et al.* 2001) but not always in premanifest (Giordani *et al.* 1995; Witjes-Ane *et al.* 2003). Manifest subjects also tend to perform badly on cancellation tasks (Bachoud-Lévi *et al.* 2001).

1.3.1.4 Language

Language comprehension appears to be relatively unaffected at least until the later stages of the disease when there is some evidence of difficulty following complex instructions (Craufurd and Snowden 2002). Problems with picture-naming are generally attributed to perceptual difficulties rather than a linguistic deficit *per se* (Hodges *et al.* 1991). In the later stages of the disease speech is affected due to problems with motor production. Bachoud-Levi *et al.* (2001) found no evidence of an impairment on the Token Test with manifest subjects although the opposite has also been reported (Lemiere *et al.* 2002) but whether this deficit was due to problems with comprehension or response selection is not discussed. The Token Test requires attention to different stimulus aspects and a series of different responses so it might be argued that an executive function, or even working memory deficit would impair performance on the task.

1.3.1.5 Memory

There is some evidence that memory declines prior to motor onset and that this decline may be less gradual than that seen in executive skills (Snowden *et al.* 2002; Solomon *et al.* 2007). Both immediate and long-term auditory recall of word lists is generally impaired in manifest HD (Giordani *et al.* 1995; Bachoud-Lévi *et al.* 2001; Lemiere *et al.* 2002) (but see also de Boo *et al.* 1997). In contrast premanifest subjects have tended not to show an impairment in these tasks (Lemiere *et al.* 2002) although a recent large study of premanifest subjects found that although they were not performing at levels indicative of clinical impairment, they were significantly below control levels (Solomon *et al.* 2007; Verny *et al.* 2007). Reports that recognition is better than recall in HD have led to theories that the deficit is primarily one of retrieval, and that this is a feature of

"subcortical" dementia. However a recent meta-analysis has challenged this, with evidence that recall and recognition impairments are of similar magnitudes (Montoya *et al.* 2006), and others have suggested that recognition is impaired in HD when false positives are taken into account (Lang *et al.* 2000a).

Digit span seems to be impaired in all but the earliest stages of HD (e.g., Lawrence *et al.* 1998b; Lemiere *et al.* 2002). Mild HD patients have shorter spatial spans than controls (Lawrence *et al.* 1996). On the same task premanifest subjects an estimated 11 years from motor onset were unimpaired relative to controls (Lawrence *et al.* 1998a) whilst a group estimated to be eight years from onset did demonstrate an impairment (Lawrence *et al.* 1998c). Memory for the source of information is also impaired in early HD in the absence of a deficit in the recall of that information (Brandt *et al.* 1995). In general memory problems in the early stages of HD are thought to reflect problems with organisation and retrieval strategies, rather than a storage deficit *per se* (Brandt 1993).

1.3.1.6 Visuospatial

Early HD patients have been shown to be impaired on perceptual matching and discrimination tasks (Lawrence *et al.* 2000) as well as face perception, drawing and block design (Craufurd and Snowden 2002), tasks which require attention to and integration of multiple stimulus parts, and some planning. Deficits are also seen in pattern and spatial recognition tasks (Lawrence *et al.* 2000). Visuoconstructional tasks such as drawing and block design, which require planning, organisation and motor skills, are also impaired; Bäckman *et al.* (1997) found that mild to moderate HD patients were impaired at the block design subtest of the WAIS as well as the copy and memory phases of the Rey-

Osterrieth figure. However, again presymptomatic subjects tend not to show impairment on these measures, although there is some evidence that this group can be impaired at visual and facial recognition (Witjes-Ane *et al.* 2003; Paulsen *et al.* 2006a).

Mild to moderate patients show little impairment on the Visual and Object Space Perception (VOSP) battery: impairments have been shown in mild patients on the object decision subtest (Lawrence *et al.* 2000) and the silhouette subtest (Lemiere *et al.* 2002) but otherwise patients tend to demonstrate intact performance, as do premanifest subjects (Lemiere *et al.* 2002). The overall pattern suggests that in the early stages of HD attentional factors may underlie perceptual problems, as basic perceptual skills seem to be preserved.

1.3.1.7 Emotion recognition

Deficits in emotion recognition have been reported in both premanifest and early HD with disgust recognition thought to be particularly impaired (Sprengelmeyer *et al.* 1996; Gray *et al.* 1997). This finding has now been replicated (Sprengelmeyer *et al.* 1997; Wang *et al.* 2003; Hennenlotter *et al.* 2004; Sprengelmeyer *et al.* 2005; Montagne *et al.* 2006), and has even been found to extend to vocal emotion recognition (Sprengelmeyer *et al.* 1996) and the olfactory and gustatory domains (Mitchell *et al.* 2005; Hayes *et al.* 2007).

However, not all studies have replicated the finding of disproportionately impaired disgust recognition in HD. For example Milders *et al.* (2003) examined facial emotion recognition in a large number of pre- and post-motor onset HD patients and found that whilst post-motor onset patients were impaired at a number of emotions, including disgust, fear was the most severely affected

emotion in the post-motor onset group and also in the pre-motor onset group. Recent data from the PREDICT-HD study have reinforced this, showing that premanifest subjects were impaired at recognition of all negative emotions (sadness, disgust, anger and fear) and showed no evidence of a disproportionate impairment in any one emotion (Johnson *et al.* 2007). The latter study also pointed out that omitting to adjust for factors such as age and IQ, which were correlated with recognition performance in their cohort, might contribute to differences between studies.

1.3.2 Longitudinal studies

Mild to moderate HD patients show more severe deficits on a wider range of tasks than early patients which suggests that cognitive performance declines as the disease progresses, and a number of studies have set out to investigate this prospectively. Longitudinally, decline in early HD has been demonstrated in some components of the UHDRS (Total Functional Capacity, Independence and Motor scores) (Bachoud-Lévi *et al.* 2001; Mahant *et al.* 2003).

In early HD patients the majority of studies report that cognitive decline can be detected in executive function and psychomotor tasks such as the SDMT, Trail Making Tasks A and B, Stroop colour and word reading, and category and phonemic fluency (Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001; Ho *et al.* 2002; Ho *et al.* 2003; Lemiere *et al.* 2004; Ward *et al.* 2006). The exception is Stroop interference in which performance only declined after taking practice effects into account in one study (Bachoud-Lévi *et al.* 2001), whilst most others find no decline even with several follow-ups over a number of years (Snowden *et al.* 2001; Lemiere *et al.* 2004; Ward *et al.* 2006). It has been suggested that this

and similar findings might be explained by a deficit in executing more automated tasks, as opposed to those under more cognitive control (Snowden *et al.* 2001).

Findings in the domain of memory are more mixed. Most studies fail to find evidence of decline in measures of auditory list learning (Bachoud-Lévi *et al.* 2001; Ho *et al.* 2003; Lemiere *et al.* 2004). Findings on digit span, spatial span and story recall are equivocal (Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001; Lemiere *et al.* 2004), whilst immediate and delayed object recall has been shown to decline but this has not been widely tested (Snowden *et al.* 2001).

Tests of language are less well represented in longitudinal studies and again findings tend to be mixed, with one study finding that Token Test performance declined over time whilst confrontation naming did not (Ho *et al.* 2003), and another the reverse pattern (Lemiere *et al.* 2004). The same two studies found either no decline, or significant decline on the VOSP. Others report no change in visual-motor integration (Ward *et al.* 2006), judgment of line orientation or mental rotation (Bachoud-Lévi *et al.* 2001) although interestingly whilst this was confirmed in a map-tracing task requiring mental rotation skills, subjects' tracing time in the same task got slower over time (a task with the same motor but fewer cognitive demands) (Snowden *et al.* 2001).

Where it has been investigated, CAG repeat length seemed to have little effect on rate of cognitive decline (Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001; Ward *et al.* 2006), whereas IQ and gender have both been reported to influence rates of decline on some tests (Snowden *et al.* 2001; Ward *et al.* 2006). The best predictor of cognitive decline in these two studies was motor ability, although a similar relationship has not been found elsewhere (Bachoud-Lévi *et al.* 2001).

Despite the fact that a number of cross-sectional studies have shown that performance decreases with estimated proximity to onset (e.g., Solomon *et al.* 2007; Stout *et al.* 2007), findings of significant decline in premanifest cohorts followed longitudinally are rare. Only one large study has found evidence of decline across executive and psychomotor tasks in premanifest subjects who were subsequently diagnosed as symptomatic during the two-year study period (Paulsen *et al.* 2001) although the same tests have been used with many other premanifest cohorts (Giordani *et al.* 1995; Lemiere *et al.* 2002; Snowden *et al.* 2002; Lemiere *et al.* 2004; Witjes-Ane *et al.* 2007).

Decline over time has also not been detected in tests of digit and visual span, verbal list learning and story recall in this population (Lemiere *et al.* 2002; Lemiere *et al.* 2004; Witjes-Ane *et al.* 2007), although object recall performance has been shown to decline around the time of clinical onset (Snowden *et al.* 2002). Examination of longitudinal change in language and visuospatial skills is again less common, and where it has been investigated there has either been no change, or for some visuospatial tests, slight improvement over time (Lemiere *et al.* 2002; Snowden *et al.* 2002; Lemiere *et al.* 2004; Witjes-Ane *et al.* 2007).

In early HD most cognitive change has been measured over a minimum of two years with few investigating it over one year or less. Time intervals in studies looking at premanifest subjects have tended to be longer, with most studies having two or three assessment sessions over at least three years.

Overall there is little decline that is detectable in the premanifest population, despite evidence that, cross-sectionally, performance at a level below controls is often shown. Decline in early manifest subjects is more notable, particularly in

the domain of executive function and psychomotor tasks, although findings on the extent and rate of change in other domains are mixed. However there are a number of factors which could influence these findings, in particular small subject numbers, different statistical methods and highly heterogeneous groups, which are discussed more in section 1.4.

1.3.3 Relationship between cognitive performance and brain atrophy

The presence of focal striatal atrophy in HD (Vonsattel *et al.* 1985; Halliday *et al.* 1998), together with findings of behavioural deficits, often leads to suggestions that striatal damage underlies whatever impairment is seen. A number of imaging studies have confirmed the existence of a relationship between striatal volumes and cognitive deficits. Caudate volume has been associated with complex psychomotor skills, verbal memory and visuospatial skills, using a CT index (Bamford *et al.* 1989); with memory/speed of processing, using volumetric MRI measurements (Starkstein *et al.* 1992; Brandt *et al.* 1995); and with score on the Mini-Mental State Examination (MMSE) (Harris *et al.* 1996). Putamen volume and metabolism have been found to correlate with neurological symptoms (Harris *et al.* 1996) and timed psychomotor and executive function tasks (Bäckman *et al.* 1997). Striatal dopamine receptor binding levels also correlate with tasks requiring scheduling and sequencing of responses (Lawrence *et al.* 1998c).

However, as discussed in section 1.2.2, it is clear that atrophy in HD is not confined to the striatum, even in the premanifest population, and there is now evidence that volume in other brain regions is associated with cognitive performance in premanifest and early HD. Measures of frontal lobe size

correlate with memory and planning performance (Bamford *et al.* 1989; Bamford *et al.* 1995; Bäckman *et al.* 1997). Insula activation has been found to be related to disgust recognition (Hennenlotter *et al.* 2004). One study included an index of whole-brain atrophy by calculating the percentage of the brain volume that was cerebrospinal fluid, and concluded that this was the best correlate of neuropsychological measures (executive, psychomotor and visual) (Harris *et al.* 1996).

Other studies have taken a whole-brain approach to this question and found regions including the thalamus (Kassubek *et al.* 2005), insula (Peinemann *et al.* 2005), white matter (Beglinger *et al.* 2005; Rosas *et al.* 2006) and widespread cortical regions (Rosas *et al.* 2005; Rosas *et al.* 2008) to be associated with cognitive performance in HD. Most of these studies focussed on a few tests in a specific domain (executive function) so it is not clear whether atrophy in similar regions is also associated with the dysfunction that is seen in other cognitive domains.

1.3.4 Summary

A number of studies have demonstrated cognitive impairments in a range of areas in patients with early and premanifest HD. Deficits have been shown in executive function (planning, shifting attention, decision-making, monitoring performance), psychomotor skills, and some perceptual and spatial skills, as well as mild memory problems, whilst language comprehension, semantic memory and skill learning and priming are relatively unaffected. Deficits on a number of tasks can sometimes be explained in terms of the executive function demands of the tasks rather than reflecting a primary problem in these areas.

Within the premanifest and early HD populations there is sometimes conflicting evidence over the type and extent of cognitive impairment. In some cases this is likely to be due to methodological differences, or lack of power to detect small effects in some studies, a fact highlighted by the evidence of deficits coming from PREDICT-HD, one of the largest studies of premanifest subjects to date (Paulsen *et al.* 2006a). However this is also likely in part to reflect the heterogeneity of the disease, and the different characteristics of the groups described as "premanifest" or "early" HD.

Overall it has been suggested that in early HD psychomotor tasks are most impaired, followed by planning, phonemic fluency, attentional span (digit span and spatial) and short-term memory (Ho *et al.* 2003). It is hard to make the same comparison for premanifest subjects but what findings there are suggest that some psychomotor tasks might decline gradually prior to motor onset, with memory undergoing a more abrupt fall-off around the time of onset (Paulsen *et al.* 2001; Snowden *et al.* 2002).

Cognitive change over time is much less clear. Again, many of the discrepancies might be due to methodological differences, some of which are discussed further in section 1.4.1. It is still unclear at what point decline in premanifest subjects can be detected and at what rate change occurs. The pattern in early HD patients is better defined but there are still many domains in which results are conflicting, and more work is needed to elucidate the pattern of cognitive progression in the disease.

Many early studies made the assumption that because the striatum was preferentially affected in HD, striatal damage must underlie the behavioural

deficits seen in the disease, and in the absence of imaging data this was not often challenged. However it is now clear that other subcortical nuclei, white matter, and cortical regions are abnormal early on in the disease and it is likely that damage to cortico-striatal networks causes at least some of the impairments seen in the disease. Future work needs to focus on clarifying the structure-function relationships in HD.

1.4 TRACKING CHANGE IN HD

Section 1.1.3 discussed the heterogeneity of HD and the fact that current clinical measures had drawbacks. However, much work is focused on developing therapies that will slow progression, if not prevent onset in HD, and good markers of disease progression are needed in order to test the efficacy of such therapies. The final section of this chapter discusses the definition of such a marker, and some of the problems associated with both cognitive and imaging measures, and the disease itself, which need to be addressed in the search for a reliable biomarker.

A biomarker is "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." (Biomarkers Definitions Working Group 2001, p91). A biomarker can be used for diagnosis or staging of disease, as an index of disease progression, or for monitoring clinical response to an intervention. In the case of HD the genetic mutation is known and this therefore serves as a diagnostic marker for the disease. However, as discussed above, few clinical measures are reliable indicators of the disease process, being highly variable, subject to floor and ceiling effects, and in some cases unable to

distinguish between symptomatic benefit and effects on disease progression. A useful marker in HD therefore needs to be sensitive to the early stages of the disease, be able to distinguish symptomatic benefit from true progression, be associated with a pathogenic process and clinical manifestation of the disease, and be able to be measured objectively and reliably; in other words it must give a clear picture of part of the disease process. It is then assumed that treatments affecting the marker will also affect the associated clinical characteristics of the disease and lead to clinical benefit to the patient. In addition, to be useful in a typical clinical trial context, a biomarker needs to be tolerated, non-invasive, reliable and reproducible in multi-centre settings.

Markers that can predict clinical benefits may potentially serve as surrogate endpoints. Currently in HD typical clinical endpoints are motor onset, and death, but trials in which these outcomes were used to assess drug efficacy would take decades. A surrogate endpoint which predicted motor onset (or death) could be used in their stead and perhaps allow therapeutic testing in the years prior to motor onset, and reduce trial times. Only a subset of biomarkers will be surrogate endpoints (Biomarkers Definitions Working Group 2001).

It is likely that no single biomarker will fully capture the extent of HD. In addition markers may not respond to treatment in predicted manners. Whole-brain atrophy rates, and rates of cognitive decline, are known to be increased in Alzheimer's disease relative to normal ageing (Jack, Jr. *et al.* 2004). It was assumed that effective treatments for the disease would reduce the rate of volume loss, as well as cognitive decline. However, a recent study of a potential treatment found that whole-brain volume loss was greater in those who had

responded to the antibody than in non-responders. In contrast cognitive change in the same subjects seemed less severe in the responders. It was suggested that the increased brain volume loss might have been caused by a reduction in amyloid deposits in the brain, rather than an increased rate of the pathological process. One effect of the treatment was therefore to invert the relationship between atrophy rate in cognitive decline in these subjects, at least over the short time course of that study (11 months). Without the concurrent measurement of cognitive performance, the improvement in clinical outcomes for the patients would not have been known (Fox *et al.* 2005). This suggests that therapeutic effects might need to be evaluated on a number of markers simultaneously in order to get a more comprehensive picture of the relationships between change in different modalities.

In HD, caudate volume has been suggested as a potential biomarker, because it can be measured objectively, is associated with the major site of pathology in the disease, and predicts motor onset (Aylward 2007). The subject numbers on which the latter statement is based are relatively small, and further work needs to be done to refine what is known about the predictive value of this measure. In addition, the work used manual measurements, and the author herself admitted that semi- or fully-automated techniques might improve the reliability of the technique. Another issue, which is relevant to all potential markers in HD, is that motor onset is itself subjective and does not fully reflect the insidious onset of symptoms across domains which is now known to occur in HD. An ideal surrogate endpoint should fully predict existing clinical outcomes, but this will depend on how well the existing clinical outcome measures reflect the disease in the first place, as well as the variability of the surrogate measure.

It is clear that more work is needed to measure the change in the natural course of HD, and to determine better indicators of onset, in order to determine which measures are the most reliable indices of the disease. This is particularly true for the premanifest and early manifest populations, the patients who are most likely to enter clinical trials.

1.4.1 Measuring change in cognition

Cognitive testing presents some difficult problems in the search for reliable markers of disease progression. Unlike more direct measures (such as MRI or rating scales like the TFC) cognitive performance can be affected by factors which are highly variable, for example subject mood or time of day. It can vary between timepoints because of these factors, and additional ones such as a change in testing venue or personnel, and so these need to be kept constant wherever possible. Like clinical scales, some tasks are subject to floor and ceiling effects, although sometimes this can partly be overcome by investigating qualitative results (e.g., analysing error patterns and strategies as well as number of correct responses) or using timed tasks. Performance is also subject to practice effects, with many subjects showing no decline, or even improvement over time, due to familiarity with the test (Bachoud-Lévi *et al.* 2001), although some tests now have alternative versions to help overcome this.

Performance often improves with increasing IQ, and declines with age. Although groups are often described as being well-matched for these variables, a non-significant difference could still affect results and so it is sensible to adjust findings for age and IQ where appropriate (see e.g., Johnson *et al.* 2007).

There are further problems with interpreting the results of other studies. Choice of control group will affect the relative impairment seen in the HD group, and this varies from healthy "normal" controls, to spouses or gene-negative siblings of HD patients, to at-risk subjects who are not told of their negative gene status until some time after cognitive testing. Different studies also use different tests to draw conclusions about the same underlying skills. Many studies which include a large battery of cognitive tests do some form of correction for multiple comparisons to limit the chance of a type I error (rejecting the null hypothesis when it is in fact true). However it is not clear what the best form of correction is (a simple Bonferroni one may not be appropriate for non-independent cognitive tests), and this also increases the chance of a type II error (accepting the null hypothesis when it is false). Given that the chosen alpha level will then also vary arbitrarily depending on the number of tests in the battery it may be preferable not to correct and simply to interpret the results with this in mind (Rothman 1990). Most published longitudinal studies of cognition in premanifest HD take this approach, or if they are more stringent still report less significant effects to give an idea of trends (e.g., Snowden *et al.* 2002). One recent study opted to correct more stringently and reported that there was no evidence of significant change in premanifest subjects (Witjes-Ane *et al.* 2007). Non-significant effect sizes and p values were not reported making it hard to compare the effects seen here with those in other studies.

In interpreting differences between studies many of these issues need to be considered as potential explanations. A number of factors can be controlled in prospective studies, for example the testing conditions or use of alternative test forms where available, whilst interpretation of results can be aided by reporting

effect sizes as well as p values. This is important both for gaining a clearer picture of cognitive progression throughout the disease and in the search for reliable markers of cognitive decline.

1.4.2 Measuring change on MRI

In comparison with measuring change in cognition, MRI measures appear to have many advantages. Brain volume is unaffected by subject mood or the scanner environment. MRI images can be analysed blind to subject identity and gene status, and if necessary data collected over a number of sites can be analysed by a single investigator and so measures can be objective and highly reproducible (Aylward 2007).

However brain volume can be affected by medication or comorbidity, as well as nutritional factors such as hydration (Duning *et al.* 2005). Although efforts are often made to exclude subjects with medication suspected of affecting brain volume, or comorbid conditions, and subjects can be advised to abstain from alcohol prior to scanning, some inter- and intra-individual variability will doubtless remain as a result of factors such as these.

Technical issues such as scanner type and consistency can also affect scan measurements. If scans are acquired over a period of time it is important to ensure that changes in scanner calibration have not biased measurements. Serial images can be registered to the same space to correct for this (Whitwell *et al.* 2004). In addition subjects need to remain still in the scanner for adequate scan quality, which means that in a disease such as HD some subjects will be unsuitable for scanning, a factor which means that cross-sectional studies will tend to focus on the premanifest or early stages, and that longitudinal studies may

risk drop-out from the more advanced subjects or only be representative of those with less severe motor problems.

Despite the fact the MRI measures are thought to be more objective and show high levels of intra- and inter-rater reliability there are again disparate findings in terms of both absolute brain volumes, patterns of atrophy, and the size of group differences in HD. In many cases these are likely to be technical and related to scan sequences or software, in which case whilst absolute measures might differ, the relative differences between groups should still be similar. Where this is not the case, analytical or statistical differences may be to blame. For example, brain volume is likely to be related to subject head size (usually unaffected by disease) and thus some sort of adjustment for this is often, although not always, made. Some researchers adjust volumes for total brain, or grey matter volume (which may be affected by disease as well as being related to head size) and hence are adjusting for the global effects of the disease, but not necessarily for head size. Semi-automated whole-brain methods often have a range of preprocessing options which can affect outcomes, and these are discussed in more detail in chapter 6.

As with cognitive findings in HD it has become clear that methodological differences in imaging studies are making it hard to generalise or to get a sense of the robustness of the results, and recently there has been a suggestion that work needs to focus on validating different methods as well as replicating earlier work (Nopoulos *et al.* 2007). The influence that different methods can have on results needs to be considered more carefully in interpreting and planning

studies, and ideally future work should consolidate disease-specific findings rather than introduce technique-specific differences.

1.4.3 Stratifying subjects with HD

HD is diagnosed when a clinician confirms the unequivocal presence of motor signs, and staging thereafter is based on the TFC score (Shoulson and Fahn 1979). Motor symptoms account for ~25% of the variance in TFC, whilst cognitive and psychiatric symptoms account for much less, suggesting that TFC may not be such a good indicator of cognitive and psychiatric problems (Nehl and Paulsen 2004). Thus subjects at the same stage of HD (including premanifest) are likely to be more similar in terms of motor scores and ability than in terms of cognitive, psychiatric or psychosocial function. This is particularly true for the premanifest / early HD distinction, since by definition this distinction must be based solely on motor signs, even if some at-risk gene carriers have been showing symptoms of cognitive or behavioural problems in the absence of motor signs. Although it is clear that as the disease advances, performance in all three domains declines, it may be that different subpopulations show "onset" and decline in domains other than motor, or that rate of decline in different domains varies (Marder *et al.* 2000; Mahant *et al.* 2003). Group analyses in HD are therefore subject to the caveat that the make-up of the groups may mask cognitive or behavioural differences whilst at the same time emphasising motor differences.

Disease duration is defined as the length of time since motor onset, and as such is subject to similar problems as staging, namely that it is predominantly an index of motor sign duration, and may not be the best measure of duration of other

symptoms. It is also somewhat imprecise, relying as it does on confirmation by a clinician, since subjects may be seen in clinic for diagnosis some time after "true" onset of motor signs.

However, within the manifest HD population the use of stages based on TFC is a way to organise subjects into approximately similar groups, and duration is a rough index of the length of time of the disease. In the premanifest population stratification based on clinical scores is more difficult since subjects tend to score at ceiling on UHDRS motor and functional assessments, and within normal ranges on cognitive tests. There have been some attempts to quantify the relative stage of the disease process in premanifest subjects based on CAG repeat length and age (Penney, Jr. *et al.* 1997) or repeat length, age and parental age at onset (e.g., Aylward *et al.* 1996). UHDRS diagnostic confidence levels for the presence or absence of motor signs have also been used (Paulsen *et al.* 2006a; Stout *et al.* 2007). More recently a probabilistic survival model has been developed which can be used to estimate probability of motor onset within a given interval, based on current age and CAG repeat length (Langbehn *et al.* 2004). The model is based on data from nearly 3000 subjects and has been widely adopted, meaning that in later work it has been possible to estimate proximity to motor onset in premanifest subjects with more confidence, and importantly that proximity to onset can be used to see if premanifest cohorts are comparable between studies. A comparison of CAG-based and diagnostic confidence level-based stratification of premanifest subjects showed that the two did not correlate, and that the former was generally more sensitive for detecting trends in more cognitive measures whilst the latter was better for motor measures (Stout *et al.* 2007).

When considering differences between studies, whatever the domain being tested, the heterogeneity of the cohorts under investigation therefore needs to be taken into account. Conflicting findings may well be the result of subjects being at different stages of the disease, and negative findings may be because small groups are highly variable, rather than because mean group differences are negligible. This is particularly true of earlier studies of premanifest subjects, some of which were carried out before an unequivocal test for the presence of the abnormal gene was available, and many of which were not able to estimate their subjects' estimated proximity to onset. One study of cognition included eight premanifest subjects with a mean age of 26 (Giordani *et al.* 1995). CAG repeat lengths were unknown but subjects would have had to have a CAG repeat length of at least 52 to be within five years of estimated onset, i.e. in the period in which deficits are more likely to be detectable. The lack of deficits cannot be taken to be representative of the whole population of premanifest HD gene carriers.

A final point when considering moving from research into clinical trials is the population that presents itself for research. Some of those who carry the HD gene and are seen in the multidisciplinary HD clinic at the National Hospital for Neurology and Neurosurgery in London do not wish to participate in research, and many who are at risk of carrying the gene opt not to have their status confirmed by testing, which precludes them from entry to a number of research studies. There is a possibility that only those with milder behavioural phenotypes, or less rapid rates of decline, enrol in and complete research studies which means that generalisation of results to the population as a whole is not guaranteed.

1.4.4 Rate of decline in HD

As mentioned above, HD is highly clinically heterogeneous and therefore it is unlikely that the disease will progress at the same rate in everyone. Two issues that are relevant to the problem of tracking change are firstly whether rate of change varies depending on disease stage or duration, i.e. whether or not change is linear; and whether rate of change varies between individuals who are otherwise at similar stages, e.g. whether genetic factors such as CAG repeat length influence rate of change.

Rate of change on the UHDRS scales has been shown to be a non-linear function of baseline scores, with minimum change when subjects scored at the low or high ends of the scales and maximum change when they were somewhere in the middle (Marder *et al.* 2000; Mahant *et al.* 2003). However this could reflect floor and ceiling effects on the scales in question as well as non-linearity in the rate of change (Mahant *et al.* 2003). Data from the large PREDICT-HD study suggest that across modalities (including volumetric imaging measures) decline in premanifest subjects starts slowly up to two decades prior to motor onset, and is then faster (but relatively linear) in the last 5-10 years before diagnosis (Paulsen *et al.* 2007), although this has been inferred from cross-sectional data and needs to be confirmed longitudinally.

A number of studies have suggested that rate of decline (functional, cognitive or volumetric) is greater in those with earlier age at onset (de la Monte *et al.* 1988; Mahant *et al.* 2003). Given that age at onset is strongly associated with CAG repeat length this has led to suggestions that rate of decline is governed by the length of the abnormal trinucleotide repeat, and this has been demonstrated for

functional, volumetric and dopamine receptor binding decline (Brandt *et al.* 1996; Aylward *et al.* 1997; Andrews *et al.* 1999) although has also been disputed (Squitieri *et al.* 2002) and is not always the case for cognitive decline (Snowden *et al.* 2001; Ward *et al.* 2006). In contrast, CAG repeat length is thought to have little effect on the extent or rate of progression of psychiatric symptoms, although this has been less widely studied (Berrios *et al.* 2001).

A recent large study has demonstrated that CAG repeat length has a small but measurable effect on rates of neurological, cognitive and functional decline, and that this effect interacts with disease duration, such that for a given duration rates are faster in those with longer CAG repeat lengths (and vice versa) (Rosenblatt *et al.* 2006). This means that change is likely to be non-linear, and different in people who otherwise appear similar (for example people of the same age or stage, but with different CAG repeat lengths).

1.4.5 Summary

Much work in HD is currently focused on finding a suitable biomarker of disease progression in order to test the efficacy of possible therapies. It is likely that more than one marker will be needed in order to give a comprehensive view of different aspects of HD. Cognitive testing and neuroimaging may provide useful measures of change although both domains have drawbacks which need to be addressed in order to maximise the reliability and reproducibility of any markers.

It is clear that many of the conflicting findings are not simply reflecting the heterogeneity of the disease. Technical or analytical differences between studies contribute to cause disparate results, as do the demographics and stratification of the HD subjects themselves.

Overall, group differences (or their absence) do not always generalise which can make it hard to draw conclusions about the profile of impairment across the disease. Future work might benefit from the use of more stringent criteria for both analysis and subject selection in order to consolidate findings.

1.5 CONCLUSION

HD is a devastating neurodegenerative disorder for which there is currently no cure, and very limited symptomatic therapy. The discovery of the causative gene has meant that it is possible to identify those who will go on to develop the disease well before the onset of overt motor symptoms and it is now known that both brain atrophy, and subtle cognitive, motor and psychiatric symptoms are detectable some time before the criteria for clinical diagnosis are reached.

Whilst atrophy is most pronounced in the striatum, a number of studies have found atrophy in cortical grey matter, white matter, and other subcortical structures, although accounts differ as to the extent of atrophy which can be seen at different disease stages, and the pattern of change. The majority of longitudinal imaging studies have focused on the striatum with much less work done on rates of change in extra-striatal regions, despite the finding that atrophy in these regions can be widespread even early in the disease.

In the early stages of HD cognitive deficits are predominantly seen in psychomotor and executive function tasks, although problems with memory and emotion recognition have also been documented. There is particular debate as to the specificity of the emotion recognition deficit in HD, and the underlying neural substrates. Language and visuospatial skills remain relatively unaffected, and problems in these areas can usually be attributed to the underlying executive

function deficits. Recent work has shown that similar deficits can be detected prior to motor onset, although effects are usually small and premanifest subjects often perform within clinically-normal ranges. A number of studies have assessed the rate of cognitive decline over time, and there is some consensus that decline can be detected reliably in some psychomotor and executive skills in early HD. Findings in the other cognitive domains are varied, and very few studies have been able to detect cognitive decline in premanifest subjects despite the fact that cross-sectional studies consistently show that performance seems to be lower in subjects who are closer to estimated motor onset.

There is some debate in the literature as to the extent to which CAG repeat length influences rate of decline; both rate of brain volume loss and cognitive decline have been associated with CAG repeat length, but not consistently. It is also unclear whether change in either domain is linear, an issue which is compounded by the fact that change on some scales may appear to be non-linear when in fact this is merely a reflection on the insensitivity of those scales at the very early and late stages of the disease. There is some evidence that change across domains is relatively linear in the years immediately prior to motor onset, but also evidence that rate of change increases with increasing time since motor onset.

With much work focusing on potential therapies for HD it is important that the natural course of the disease is well-documented, and in particular that reliable measures of disease progression are identified in order to have something against which to test the efficacy of such therapies. The fact that many findings in HD are equivocal has led to some researchers to call for more careful stratification of subjects, and more validation of different neuroimaging techniques, in order to

consolidate what is known about the disease (Nopoulos *et al.* 2007; Witjes-Ane *et al.* 2007).

PROBLEMS THAT NEED TO BE ADDRESSED IN HD

Huntington's disease is an autosomal dominant neurodegenerative disease, usually of adult onset, for which there is currently no cure. Since the discovery of the causative gene (Huntington's Disease Collaborative Research Group 1993) it has been possible to identify gene carriers prior to symptom onset and therefore to learn more about the disease in its very early stages. Since then much work has focused on identifying markers which could predict disease onset or monitor disease progression, in order to test the efficacy of potential therapies. Both neuroimaging and cognitive measures have been used to demonstrate the extent of brain atrophy and associated functional deficits in the disease, and to try and track change over time.

However, many findings in HD are equivocal, making it difficult to generalise results and to feel confident about the robustness of findings. This is partly due to the clinical heterogeneity of the disease and the fact that stratification of patients with HD into relatively homogeneous groups is difficult. Some discrepant findings can also be attributed to very different uses of imaging techniques, and to inadequate consideration being given to the influence of non-disease-specific factors on results. The natural course of the disease needs to be more carefully documented if reliable markers of disease progression are to be found.

AIMS OF THIS THESIS

This work aims both to disambiguate some of the findings discussed above, and to investigate the pattern of change in HD with a view to suggesting more reliable measures of disease progression. It also aims to address a number of methodological issues highlighted previously, by using a large, well-defined cohort, optimising image analysis techniques, and taking care to control for non-disease-specific confounds in analysis. Both neuroimaging and cognitive measures will be investigated in order to define the pattern of neurodegeneration and associated cognitive changes in HD better.

Specific aims are to:

1. assess the pattern of atrophy of the HD brain using both manual volumetry and voxel-based morphometric methods;

2. identify relative deficits across a broad range of cognitive domains, and investigate the association between cognitive and neuroimaging measures, with particular focus on emotion recognition and its neural correlates;

3. compare the ability of cognitive and neuroimaging measures to detect change over time in HD, and investigate whether CAG repeat length is associated with rate of change;

4. assess a semi-automated measure of whole-brain atrophy as a potential marker of disease progression in terms of whether it can detect disease-related change and whether this is predictive of clinical function.

2 GENERAL METHODS

2.1 SUBJECTS

2.1.1 HD gene carriers

The majority of subjects described in this thesis were recruited from the multi-disciplinary Huntington's disease clinic based at the National Hospital for Neurology and Neurosurgery, Queen Square, London. Further subjects were recruited from the Huntington's disease clinic run by Dr Roger Barker at Addenbrooke's Hospital, Cambridge. Subjects were diagnosed as carrying the HD gene following a DNA test. Subjects were included with a CAG repeat length of 40 or higher (Rubinsztein *et al.* 1996). Only those subjects who had had a positive test for the HD gene were included in these studies; subjects with an affected parent who were at-risk of developing HD but who had not been tested were not recruited.

2.1.2 Healthy controls

Neurologically healthy subjects were recruited as controls. Most of the controls were spouses or partners of HD gene carriers who were in the study, or were non-gene carriers from an at-risk family (i.e. they had had a negative test for the HD gene). Some controls in the pilot study (see chapter 3) were healthy volunteers from the National Hospital or University College London. Controls in the main longitudinal study (chapters 4 – 11) were offered £50 over the course of the two study visits, although most declined this and requested that the money be put towards research costs.

2.1.3 Exclusion criteria

Exclusion criteria for all subjects were contraindication to scanning (such as metal implants or claustrophobia), history of substance abuse, or significant medical, neurological or psychiatric comorbidity, i.e. conditions that were non-HD related. Detailed records were made of medications and supplements. No subjects were taking any medication known or suspected to influence brain volume.

2.1.4 Ethical issues

All subjects were offered reimbursement for all travel and meal costs for themselves and a companion. All subjects gave written informed consent for involvement in all the studies, which had Local Ethics Research Committee approval. Patients were followed up clinically at either the London or the Cambridge clinic. Research assessments had the potential to increase anxiety about the implications of HD for the patients, and so patients were made aware that if they were concerned about any aspect of the study or their performance in it they could be referred to the appropriate clinical specialist for further assessment, advice and support.

2.2 CLINICAL ASSESSMENT

All subjects were assessed by a neurologist who took a full medical and family history. Subjects were also assessed using the clinical part of the Unified Huntington's Disease Rating Scale (UHDRS) (Huntington Study Group 1996) which comprises:

1. Standardised ratings of the typical motor features seen in HD, leading to a motor score ranging from zero (normal) to 124.

2. Functional assessments: a series of yes / no questions about the subject's ability to perform everyday functions, and a Total Functional Capacity (TFC) (Shoulson 1981) score based on the subject's capacity to work, manage finances, do domestic chores, activities of daily living, and care requirements, which yields a score ranging from 13 (normal) to 0. It is from this score that clinical stage is derived (Shoulson and Fahn 1979), 11-13 being stage 1, 7-10 stage 2, 3-6 stage 3, 1-2 stage 4 and 0 stage 5.
3. An independence scale against which the subject's current level of independence is judged, ranging from 100 ("No special care needed") to 10 ("Tube fed, total bed care").
4. Standardised ratings of the severity and frequency of behavioural symptoms.

Patients (but not controls) in the pilot study (chapter 3) were assessed using the full UHDRS rating scale. All subjects in the other studies were assessed using the first three components of the scale mentioned above, and a novel behavioural assessment scale, the Short Behavioural Assessment (SBA), a short version of the Problem Behaviours Assessment for HD (PBA-HD) (Craufurd *et al.* 2001). Subjects in the latter studies also filled out a Beck Depression Inventory (BDI-IA) (Beck and Steer 1993). The UHDRS also includes a set of cognitive tests, which for all subjects in this thesis were administered as part of the neuropsychological battery (described in the relevant chapters).

2.3 DNA TESTING

All premanifest or early HD subjects had undergone a DNA test prior to inclusion in the study (see section 2.1). Subjects carrying the HD gene in studies other than the pilot study (chapter 3) consented to a further DNA test and all

CAG repeat lengths for these patients were subsequently resized in the same laboratory.

2.4 COGNITIVE ASSESSMENT

All subjects underwent neuropsychological assessment, with batteries designed to be comprehensive but brief enough to administer in a single session. All assessments were done by me. Wherever possible subjects were tested in a single session, during which they were offered refreshment and breaks as needed. Details of each battery are given in the relevant chapters, and descriptions of the standard neuropsychological tests are in Appendix 3.

2.5 IMAGING

2.5.1 Imaging as a measure

MRI has become the imaging modality of choice for the brain with its multiple different parameters that can be altered to achieve different contrasts between tissue types, and true three-dimensional volumes have become standard. However, because these images are a visual reconstruction of the radio signal emitted by different tissue types in the MR scanner it is important to realise that there is not a one-to-one correspondence between the brain and the MR image; this depends on the tissue qualities and scan parameters (Figure 2-1).

The intensities of the voxels that make up the image vary depending on the tissue type which they represent and make it possible to visualise brain structure *in vivo*. Software can be used to outline structures of interest in the image, either manually (based on visual judgement of the border between two tissue types) or semi-automatically (e.g. by setting an intensity threshold such that voxels below a certain cut-off are excluded, leaving only those voxels which represent the

tissue type of interest) (see section 2.5.3.1). Since each voxel represents the signal from a particular 3D region in the brain it is possible to sum the voxels contained within a structure outlined in this manner and hence calculate its volume.

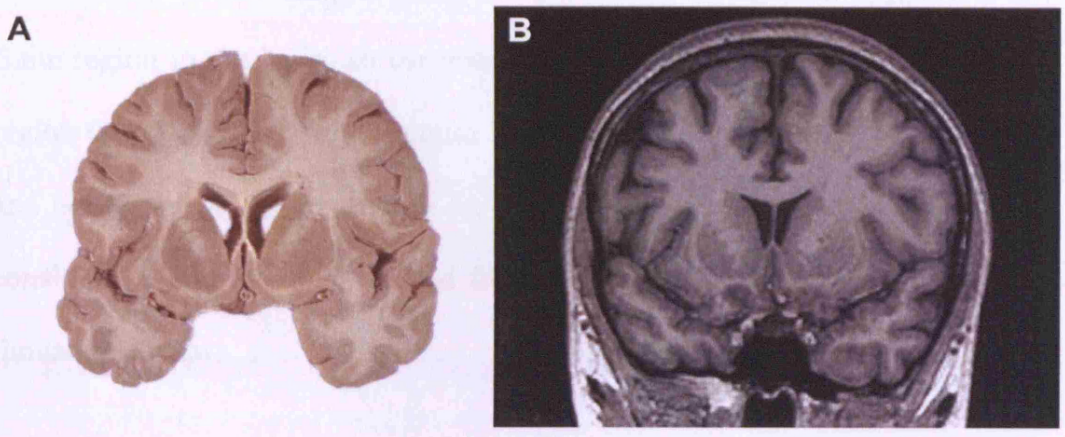


Figure 2-1 A) shows a post-mortem brain in coronal section, and B) a similar slice from a brain imaged *in vivo* using T1-weighted MRI
Note that there are regions in the MR image where the grey / white contrast and tissue boundaries are not very well-defined compared with the post-mortem section

Differences in intensities between corresponding voxels on one or more scans can also be used to estimate inter-subject differences, or intra-subject change over time. For example, the technique of voxel-based morphometry (see sections 2.5.3.2 and 2.5.5) uses differences in the normalised intensities at each voxel in a set of images to measure group differences in volume at that voxel. Alternatively, if two scans from the same subject, taken some time apart, are aligned and displayed in the same space, differences in intensity between corresponding voxels can be used to estimate volume loss, or change over time.

All these measures are based on the intensity at each voxel, which as well as representing signal from brain tissue, will contain some noise caused by factors

such as variation in the scanner (e.g., in the magnetic field strength), movement of the subject, the positioning of the subject in the scanner, or the particular software used to reconstruct the signal. Some of these sources of error can be measured and adjusted for (see section 2.5.4.1). In addition, there are limits to the resolution of the images, both in the sense that the signal is acquired from a finite region in the brain, so the image cannot represent differences within that region (pixel or voxel), and because some tissue types may emit similar signals and hence not be distinguishable on certain types of image. It is important to consider that any measure derived from MRI will only be as accurate as these limitations allow.

The following sections describe in more detail image acquisition, the two software packages used for estimating volumes and group differences, and the image processing involved.

2.5.2 Acquisition

All images were acquired on a 1.5T Signa MRI scanner (General Electric, Milwaukee, Wisconsin, USA) using a spoiled gradient echo technique. Scans generally included a sagittal T1-weighted localising sequence, and a T1-weighted volumetric image. At the first visit scans usually also included an axial dual-echo sequence (T2-weighted and proton-density weighted). The dual echo sequence was checked by an experienced radiologist for non-HD pathology and not used further in any analysis. Exact imaging parameters for each study are given in Appendix 1.

Volumetric images were reconstructed as 124 contiguous 1.5mm coronal slices and images were digitised for analysis.

2.5.3 *Software*

2.5.3.1 *MIDAS*

The Medical Information Display and Analysis System (MIDAS) software was used for the volumetry described in this thesis. MIDAS was developed by Freeborough *et al.* (Freeborough *et al.* 1997) and runs on both Unix and Linux platforms and is implemented in the C programming language. The software allows 3D data to be displayed simultaneously in two or three orthogonal views (see Figure 2-2).

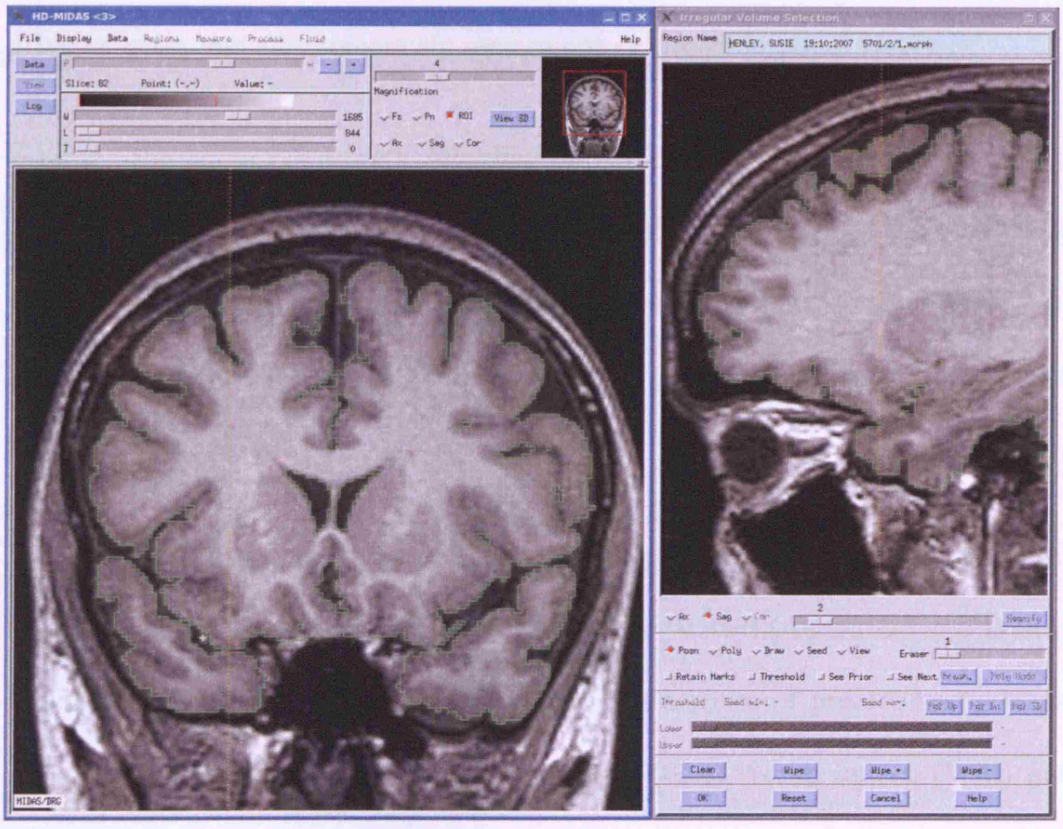


Figure 2-2 Screenshot of MIDAS software showing a whole brain being outlined in coronal and sagittal views

Brain structures can be outlined using both semi-automated and manual techniques. Manual techniques rely on the operator delineating borders with a mouse-driven cursor, based on visual judgement of intensities. Semi-automated

techniques usually involve a threshold being set by the operator which then excludes all voxels above or below a certain intensity. Segmenting brain from non-brain in MIDAS is semi-automated and uses interactive thresholding to generate an initial region. This region can then be eroded and dilated a number of times, to separate brain tissue from non-brain (see Appendix 2). Segmentation of sub-structures such as the caudate nucleus is also semi-automated; thresholds which separate the structure of interest from surrounding tissue are used to delineate borders automatically, after manual placement of a seed within the structure. The borders can then be edited manually in one view using the mouse-driven cursor, and simultaneously seen in one or two orthogonal views.

2.5.3.2 Statistical Parametric Mapping (SPM)

SPM is an image analysis package developed by Friston *et al.* at the Functional Imaging Laboratory, Institute of Neurology, London (Friston *et al.* 1995). The work in this thesis used SPM2 running on a Matlab platform (MathWorks, Natick, Mass., USA).

When SPM is used to analyse structural (rather than functional) images this technique is known as voxel-based morphometry (VBM) (Ashburner and Friston 2000). VBM uses the SPM package to perform a series of preprocessing steps, including segmenting images into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), and spatially aligning these segments into standard sterotaxic space (see section 2.5.5.1). After the segments are aligned a statistical test is performed at each voxel. In general intensity at each voxel (which represents either volume or tissue density) is regressed on a number of explanatory (or nuisance) variables to look for differences between groups or

associations between variables. Whether the voxel represents volume or tissue density depends on the preprocessing and is discussed in more detail in section 7.4.3. The resulting map of the brain highlights all those voxels for which the differences or associations are statistically significant.

2.5.4 Image Processing

2.5.4.1 Intensity inhomogeneity correction

Ideally, scans will have good contrast between tissue types (grey matter, white matter, and CSF), and uniformity within a tissue type. This makes structural boundaries clear and aids both manual and semi-automated (interactive thresholding) segmentation, and registration methods which use intensity-based matching algorithms. However scans can be subject to intensity inhomogeneity, or bias, (for example, the mean intensity varies smoothly from one side of the image to the other), for a number of reasons, including inhomogeneity of the magnetic field in the scanner, inhomogeneity of the radiofrequency pulse or nonuniformity in the sensitivity of the receiver coils used to detect the MR signal (see, for example, Vovk *et al.* 2007). Prior to any analysis images in this study had such intensity inhomogeneities corrected using the N3 algorithm (Sled *et al.* 1998).

However if such fields are not entirely removed this can also cause problems when scans are registered into the same space for comparison (see section 2.5.4.2). If any residual bias field on the two images is different this will introduce an artificial difference in the intensity value of corresponding points on those images, which could affect intensity-based atrophy calculations (Lewis and Fox 2004). In the work presented in this thesis the differential bias field between

registered scan pairs was estimated and adjusted for; this does not remove bias *per se* but removes differences in the bias fields so that the residual inhomogeneity in each scan is similar, thus reducing differences in intensity which were a product of the different bias fields (Lewis and Fox 2004).

2.5.4.2 Linear registration

Serial MRI is useful to detect brain volume change over time. Change in volume between two timepoints can be derived by subtracting the volume of the region of interest on a repeat scan from that on a baseline scan. However the resulting difference score is susceptible to error, in particular from segmentor inconsistencies or from scanner drift (changes in scanner calibration over time which could mean for example that voxel size changes slightly over time).

Registration of serial MRI scans allows more precise quantification of volume changes, and also allows regions of change to be visualised. In simple terms, the repeat image is spatially aligned to the baseline, and the alignment process can correct for problems such as change in voxel size (Freeborough *et al.* 1996; Whitwell *et al.* 2004) and allow a direct visual comparison of the scan pair from which the difference in volumes is measured. More specifically, the registration algorithm aims to align a follow-up image (FU) to match a baseline image (B). This is an optimisation process, in which a set of parameters that will transform FU to B are calculated that will maximise a similarity measure, or cost function. After finding the best set of parameters this transformation is applied to FU which is resampled to have the same image space as B. As voxels in FU are unlikely to map exactly onto voxels in B an interpolation model is used to work out the correct voxel value; the voxel value in the transformed image is some

function of the values of neighbouring voxels in the original FU image. In practice, interpolation is also needed during estimation of the transformation parameters in order to calculate the cost function.

In linear registration the transformation parameters are applied equally to every voxel in the image. Once the two images are aligned a point in one image should correspond to the same point in the other. In practice, where volume has been lost over time the two images will not correspond exactly and the amount by which they differ, which will often be seen around borders, can be used to estimate the amount of volume lost (see Figure 2-3).

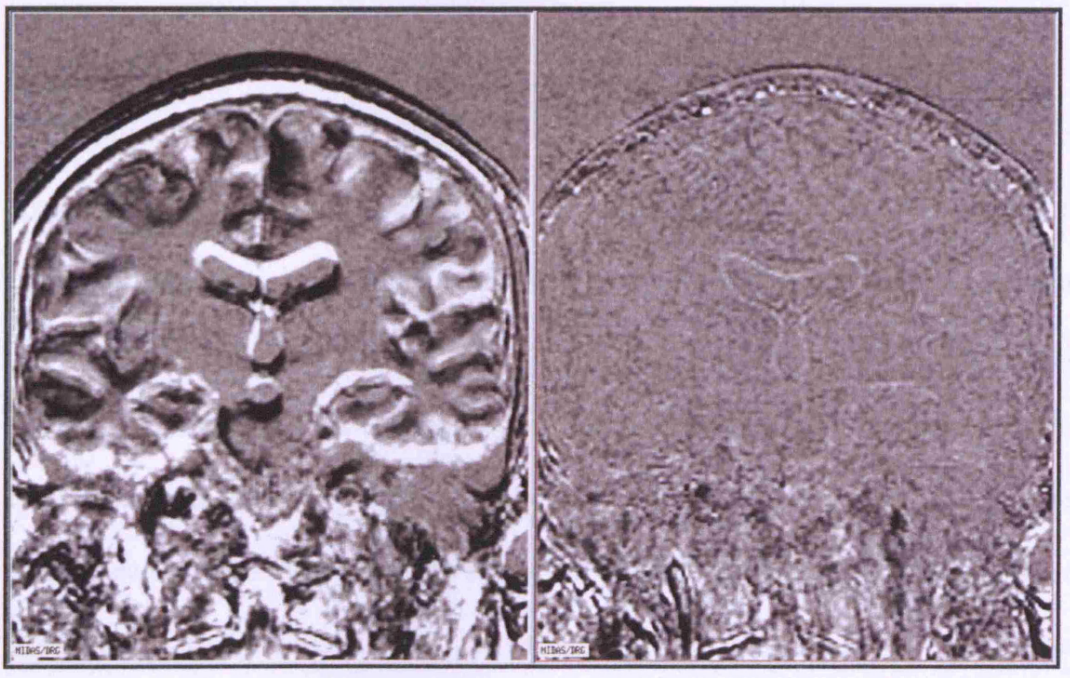


Figure 2-3 Left panel shows a difference image derived from two non-registered scans taken a year apart; right panel shows a difference image from the same scans after 12dof registration

The left difference image highlights the misalignment of the scans, whilst the right highlights regions that remain different after realignment, most obviously around the ventricles. Note that the skull, neck and ears are less well-matched, because the registration algorithm is applied to the brain region only.

The work in this thesis used Automated Image Registration (AIR, v 5.2.5) (Woods *et al.* 1998). Registrations were affine (12 degrees of freedom, 12dof) which meant that the transformations were translations, rotations, scalings or shears along each of the three orthogonal axes of the image. The cost function to be minimised was the standard deviation of the ratio of the intensities of the two images (calculated at each voxel). This is a measure of the uniformity of the ratio image; the more uniform the ratio image, the more likely the two images are to be well-aligned. Follow-up images were resampled using chirp-z interpolation (Rabiner *et al.* 1969). Affine registration is also implemented through MIDAS (section 2.5.3.1).

2.5.4.3 The Boundary Shift Integral

Registration corrects for changes in voxel size due to adjustment or drift in scanner calibration over time (scanner drift) and therefore measuring the difference between the volumes of two registered images would improve upon a difference measure based on unregistered images. However this method still relies on the accuracy of the two segmentations. Once two images have been registered it is possible to estimate volume change by measuring directly the difference between the boundaries of the two aligned structures of interest, rather than measuring the volume of each separately. This makes the measurement less susceptible to segmentor error than simply subtracting the segmented volume of the follow-up image from that of the baseline image.

The boundary shift integral (BSI) uses the shift in intensity differences on a pair of registered intensity-normalised images (i.e., where the edges of the structures have moved from baseline to repeat scan) to quantify the difference between the

two (Freeborough and Fox 1997). Firstly a 3D region within which this difference is to be measured is defined. The regions of the structure of interest (for example, the whole brain) are outlined on both scans using MIDAS (see section 2.5.3).

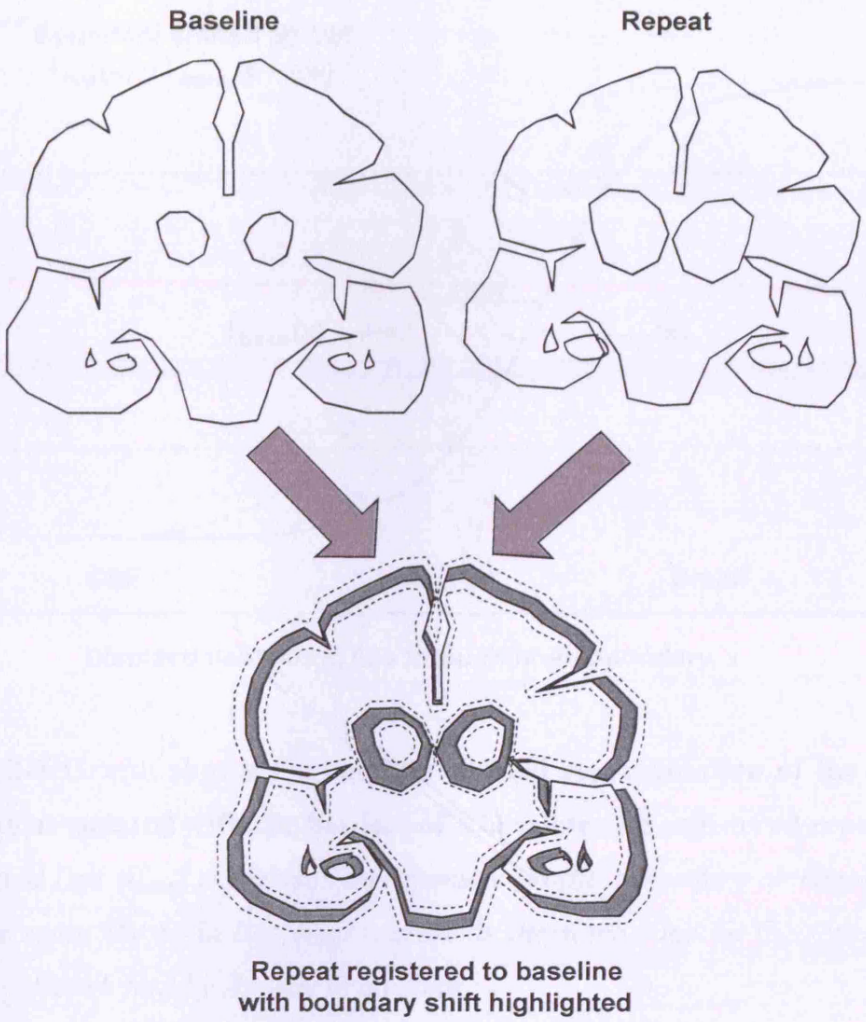


Figure 2-4 Schematic showing repeat brain registered to baseline and the resulting shift in boundary

The shift in boundary between the two images is highlighted in grey. The boundary shift integral is calculated in a region encompassing this shift: the intersection of the two brain regions is eroded by one voxel to give an inner boundary (inner dotted line), and the union of the two regions is dilated by one voxel to give an outer boundary (outer dotted line). Together these define a region which encompasses all the borders on the two images.

These regions are used to define the region in which the boundary shift integral is calculated (see Figure 2-4). Figure 2-5 shows a plot of the changes in intensity that might be expected as one moves across the CSF/GM border of a brain image.

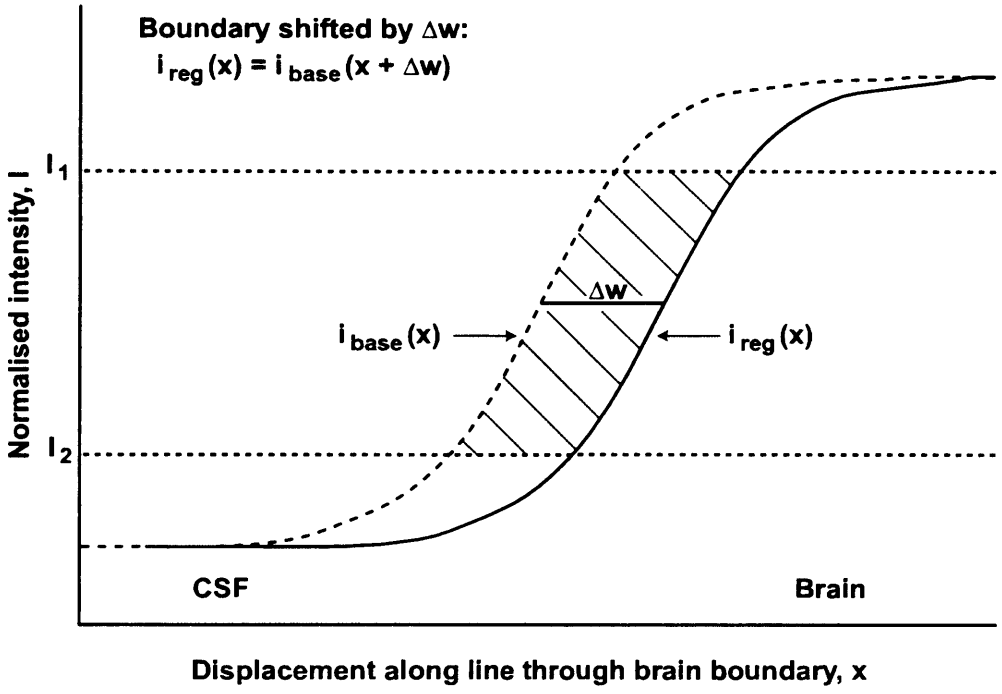


Figure 2-5 Graph showing a one-dimensional representation of the shift in intensity associated with the borders of a baseline and registered repeat scan
The dotted line (i_{base}) represents the intensity profile (boundary or edge) for the baseline scan; the solid line (i_{reg}) represents the same edge on the repeat scan; that edge has shifted by Δw due to atrophy.

We wish to know Δw, the distance through which the boundary has shifted along x. A simple measure would be to quantify the shift at a single value of I (where I is an intermediate intensity, between that inside the brain, and that outside). However Δw would then be dependent on the value of I that was chosen, so a more robust estimate is to estimate Δw over a range of values of I, and then average them (where the range of values of I falls between the intensity in the

bulk of the structure, and the intensity on the other side of the boundary). This is the integral, with respect to I, of the area between the curves within the range of I chosen (equation 2.1).

$$\Delta w = \frac{1}{I_1 - I_2} \int_{I_1}^{I_2} (x_{base}(i) - x_{reg}(i)) di \qquad (2.1)$$

In practice however we do not measure the spatial distance between two points of equal intensity, but the difference in intensity between two spatially aligned points: this is represented as the integral with respect to x.

This integral can be extended to three-dimensions to measure change in volume. Δv. The integral is estimated by sampling the image at small intervals (every voxel) and multiplying the result by voxel volume. The integral is only evaluated within the set of voxels already defined (see above, and Figure 2-4) as encompassing the boundary regions of both images.

The range of the integral is the intensity window, which is defined in terms of its centre $I_c = I_1/2 + I_2/2$, and width $I_w = I_1 - I_2$. Ideally this window should fall within all the intensity transitions that are associated with the boundaries of the structure of interest. However in order to improve the signal to noise ratio the window should be made as wide as possible. In practice the optimal window can be determined as that which maximises the agreement between the volume difference as measured by the BSI, and the volume difference as measured by segmenting the two images. Ideally this optimisation process should be run on simulated data (where the real change in volume is known). Where this is not possible the optimal parameters can be determined from a subset of data and then applied to the remaining images.

Atrophy rates measured using the BBSI show relatively good agreement with rates calculated by volume subtraction, although atrophy may be consistently underestimated very slightly because of scan variability (e.g., in a noisy scan the boundary region might fall slightly outside the chosen intensity window). However, within-group variance (which may be attributable to scanner or segmentor error) is reduced and separation between groups increased (Freeborough and Fox 1997). This technique has been well validated in Alzheimer's disease (AD) and normal ageing and shown to have a high level of precision for measuring cerebral changes including rate of whole-brain atrophy (Fox and Freeborough 1997; Scahill *et al.* 2003).

2.5.5 Voxel-based morphometry

The images were processed using MATLAB 7.0 (The MathWorks, Inc, Natick, Massachusetts, USA) and SPM2 (Wellcome Department of Cognitive Neurology, ION, London). VBM was performed following a modified version of the optimised method (Good *et al.* 2001b). The background to this method and its suitability for the cohort used here is the subject of section 6.1.1.

2.5.5.1 Image preprocessing

The preprocessing steps are represented in Figure 2-6. The initial processing used a modified version of the script provided by Christian Gaser (available at http://dbm.neuro.uni-jena.de/vbm/). The native space images were affine-registered using the standard SPM2 T1 template to put them into standard space (initial normalisation), and underwent initial segmentation.

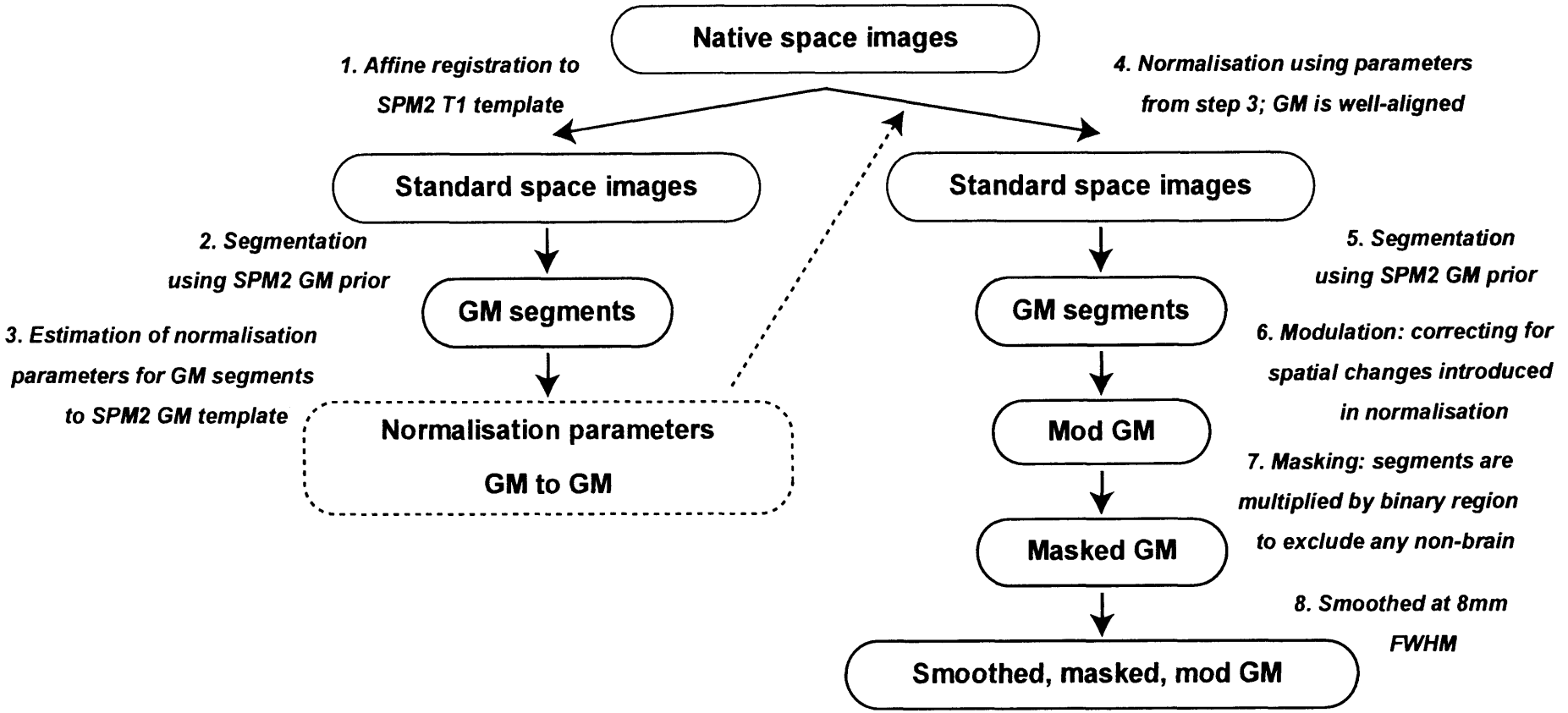


Figure 2-6 VBM preprocessing

The native space images are affine-registered and segmented, and the grey matter segments are used to estimate normalisation parameters to the grey matter template (left side). These normalisation parameters are then used to normalise the native space images again, and these are segmented, modulated, masked and smoothed prior to statistical analysis.

New normalisation parameters were estimated for warping the resulting grey matter segments onto the SPM2 grey matter template. These normalisation parameters were then used to warp the original native space images.

SPM2 uses a combination of prior probability maps and intensity histograms to segment these images automatically into GM, WM and CSF, after which voxel values range from 0-1 and represent the probability of each voxel being the particular tissue type. Because the normalisation process could correct for some volume differences which were actually due to atrophy, the segmentations were modulated. Modulation entails multiplying the value of each voxel by the Jacobian value from the normalisation process. The Jacobian is an index of by how much the voxel was expanded or contracted during normalisation so after the modulation process voxel intensity represents volume.

The WM and CSF segments were not used in any further analysis as this work was primarily focused on grey matter changes. Preliminary work (discussed in detail in section 6.1.1) showed that normalising grey matter to a grey matter template resulted in better alignment of the grey matter segments compared with standard normalisation. This was achieved with the standard SPM templates; using customised (study-specific) templates for normalisation and segmentation did not produce any additional improvement. Each grey matter segment then had spurious non-brain tissue removed according to a brain mask derived from the corresponding original image using MIDAS (Freeborough *et al.* 1997). Finally the images were smoothed using an 8mm full width half-maximum (FWHM) Gaussian kernel.

2.5.5.2 Presentation of statistical parametric maps

After models were fitted and parameters estimated (described in the relevant chapters) an explicit mask was applied to the maps to exclude any voxels for which more than 10% of the images had a value of less than 0.1. This "majority masking" was preferred to the default "absolute" mask option in SPM, which would exclude any voxels for which one or more images had a value of less than 0.1 and thus perhaps be unduly influenced by a single poorly-registered or highly atrophied scan. All maps were displayed as overlays on a smoothed version of the Montreal Neurological Institute (MNI) 152 T1 template. MNI coordinates were converted to Talairach coordinates using the mni2tal MATLAB function provided by Matthew Brett (http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach) and regions of atrophy were localised using the Talairach daemon developed by Lancaster *et al.* (http://ric.uthscsa.edu/new/resources/talairachdaemon/talairachdaemon.html).

A statistical parametric map consists of many comparisons (one at each voxel) and so some form of correction is usually applied (the threshold for what is held to be statistically significant is adjusted to decrease the chance of false positive results). However, points in the image are spatially correlated to some degree, both because of the inherent properties of the image, and because of the smoothing applied during preprocessing, so a simple adjustment for independent comparisons (e.g. Bonferroni) is not appropriate. The degree of correlation can be estimated by estimating the smoothness of the image, using the residuals from the fitted model, and the smoothness in each direction is multiplied to give the size of a "resolution element" or resel. The number of resels is an estimate of the number of truly independent samples in the image and once it is known it can be

used to calculate a threshold for the t map, at a given alpha (Worsley *et al.* 1992). With family-wise error (FWE) correction (see chapter 6) the threshold depends on the smoothness of the image, and the number of resels, as well as the underlying effects, and hence smoothness and resel counts are reported for each analysis in this thesis.

2.6 STATISTICAL ANALYSIS

Statistical analysis of volumetric, cognitive and clinical data was done using STATA version 9.2 (Stata Corporation, College Station, Texas, 2006). Statistical analysis of structural brain images on a voxel-wise basis was performed using SPM2 (see section 2.5.3.2).

In general significance levels were not adjusted to take into account the number of comparisons because all the associations being investigated were thought to be of independent scientific interest (Rothman 1990). One of the primary aims of the thesis was to compare a broad number of tests as measures of change over time and false acceptance of the null hypothesis might mean that many potentially sensitive markers were discarded.

Significance levels were however adjusted for the VBM analyses, as in this case many thousands of comparisons were being made simultaneously and they were not of independent scientific interest.

3 USE OF SERIAL MRI AND COGNITIVE TESTING TO TRACK CHANGE IN HUNTINGTON'S DISEASE: A PILOT STUDY

3.1 INTRODUCTION

As discussed in chapter 1, HD is clinically heterogeneous and tracking change (disease progression) over time reliably has proved challenging using clinical, cognitive or imaging measures. This chapter describes a small study which was run in order to pilot the BBSI (see section 2.5.4.3), which had not been used in HD before, as well as to gather cognitive and clinical data on the same cohort.

There is increasing evidence that atrophy in HD occurs throughout the brain, even in the early stages (see section 1.2.2). Whole-brain atrophy, although it does not provide information about the exact location of loss, can be measured using semi-automated methods which are quick and tend to be less susceptible to segmentor error than manual outlining (Freeborough and Fox 1997). The brain boundary shift integral (BBSI), which estimates whole-brain volume loss from the intensity differences around the borders of a registered pair of scans, is one such method which has already been used in clinical trials in Alzheimer's disease (Fox *et al.* 2005).

In the early stages of HD, whole-brain atrophy may well be relatively small and within normal ageing ranges if atrophy is essentially confined to the striatum, which loses about 0.5ml per year (i.e. less than 0.1% of total brain volume per year) (Aylward *et al.* 2000). One of the aims of the pilot study was therefore to determine whether there was enough signal (i.e. diffuse losses), over relatively short time periods, for increased whole-brain atrophy rates to be detected with this technique. Although at autopsy atrophy is marked throughout cortical

regions it is unclear when atrophy spreads beyond the striatum (Halliday *et al.* 1998). In addition, one of the signs of HD is chorea, and movement during an MRI scan can affect its quality and hence the accuracy of any measure derived from it. Therefore this study also set out to investigate whether subjects could tolerate repeat MRI scans well enough to obtain good quality serial imaging so that change over time could be measured accurately.

Cognitive findings in HD have also been mixed, and there is a lack of consensus over which skills show reliable, measurable decline in the early stages (Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001; Ho *et al.* 2003). Longitudinal cognitive testing is also hampered by the existence of practice effects, as well as environmental factors which are hard to control at different timepoints, such as subject mood and fatigue, and testing conditions. Therefore a second aim was to pilot a fairly comprehensive battery of tasks, covering areas which are known to be affected in HD as well as those which are thought to be relatively spared, in order to see which tasks were most useful for measuring decline, given the problems mentioned above; these findings would then be used to refine the battery for future studies.

Few studies in HD have measured change over intervals shorter than one year, or combined structural imaging measures with a comprehensive battery of psychological tests. Some that do have used CT measures or reduced large cognitive batteries to a number of factors so that conclusions about structure-function relationship remain limited (e.g., Starkstein *et al.* 1992; Bamford *et al.* 1995). Therefore the pilot data were also used to assess the relationship between whole-brain atrophy, and cognitive and clinical decline. Initially the study aimed

to assess subjects twice, a year apart, but an unforeseen scanner software upgrade was then planned during month nine of the study. Since such an upgrade could affect scan quality and hence the BBSI measure, the second assessment was brought forward to six months, with a further cognitive and clinical follow-up at twelve months, so that both baseline and six-month scans were obtained before the software change.

In summary, this chapter describes a pilot study, set up to examine the tolerability of serial MRI and usefulness of the BBSI as a measure in HD, and to investigate how cognitive measures changed in the same cohort, with a view to introducing the best measures into a larger follow-up cohort, if successful.

3.2 METHODS

3.2.1 Subjects

Fifteen subjects with genetically-confirmed HD were studied (all early disease at baseline with 10 subjects at stage 1, and five at stage 2 (Shoulson and Fahn 1979)). Nine controls were group-matched for age, sex and premorbid IQ. As subjects with early HD displayed motor symptoms it was not possible for clinical or neuropsychological examinations to be performed blinded to gene status; however scan segmentation was done blind to gene status.

3.2.2 Assessments

3.2.2.1 Clinical assessment

Subjects were clinically assessed at baseline and 12 months by the same two senior neurologists whose intra- and inter-rater reliability coefficients were >0.9,

using the UHDRS clinical rating scale which comprised motor, psychiatric and functional assessments (Huntington Study Group 1996) (see section 2.2).

3.2.2.2 MRI acquisition

Subjects had T1-weighted volumetric inversion recovery prepared, spoiled gradient recall scans on the same 1.5T GE scanner at baseline and six months (for parameters see Appendix 1). Whole brains were segmented as described in Appendix 2, with the segmentor blinded to group membership. Six-month scans were positionally matched to their baseline image using an affine registration with 12 degrees of freedom, and change in brain volume was quantified from registered pairs using the BBSI (Freeborough and Fox 1997) (see sections 2.5.4.2 and 2.5.4.3).

Registered scan pairs were checked by an experienced imager with no knowledge of disease status, and any pairs which were inadequate due to motion or other artefact were discarded from the analysis. Only those subjects who had adequate MRI follow-up, 13 subjects and seven controls, were included in the longitudinal analysis. Demographic details of these subjects are given in Table 3-1.

3.2.2.3 Cognitive assessment

All subjects underwent neuropsychological assessment at baseline, six months and 12 months. The battery was designed to be comprehensive but brief enough to complete in a single session (typical duration was under two hours). Whilst it focussed on areas known to be impaired in HD, e.g. executive function, psychomotor and memory tasks, it also covered naming and visuoperceptual skills, which were predicted to be less affected.

The battery comprised the following tests, further details of which are in Appendix 3:

Estimated Pre-morbid IQ: National Adult Reading Test (NART). As performance on other cognitive tasks can be IQ-dependent it was important to obtain an estimate of pre-morbid IQ so that this could be controlled for in the analysis.

Executive function / psychomotor tasks: three tasks were included as part of the standard UHDRS assessment. They were the Stroop word-reading and colour-word interference (old version) (Stroop 1935); phonemic (letter) fluency (Benton and Hamsher 1978); and the Symbol Digit Modalities Test (SDMT) (Smith 1968). These tests have a motor component but are also thought to test more "executive" skills; word generation is thought to depend to some extent on frontal lobe integrity, as is the interference task of the Stroop in which subjects must ignore the automatic reading response in order to name the ink colour.

The trail-making subtest of the D-KEFS (Delis *et al.* 2001) which has separate sections for number cancellation, number sequencing, letter sequencing, alternating sequencing, and motor speed, was included. This task allows one to measure scanning time, motor speed, and basic sequencing, before the more complex number-letter sequencing test.

A simple reaction time task was designed and presented on SuperLab Pro 2.0 (Cedrus Corporation, San Pedro, CA). This required subjects to use the index finger of their dominant hand to press one of four numbered buttons on a response pad. The numbers 1-4 were centrally presented on the screen in a

pseudo-random order in 20 blocks of 4, with a pause after the 10th block. Subjects were requested to press the buttons as quickly as possible and mean reaction time was measured.

Memory: immediate memory was assessed using digit span forward and backward with three trials at each length (Wechsler 1981). Three trials were used rather than the standard two because this tends to give a more robust measure of the subject's span. The Recognition Memory Test (RMT) (Warrington 1984) was also included, assessing face and word memory.

Naming: naming was measured using the Graded Naming Test (GNT) (McKenna and Warrington 1983). This was presented on a computer using SuperLab Pro 2.0 (Cedrus Corporation, San Pedro, California, USA). Stimulus presentation was controlled by the investigator and response latency was recorded by experimenter button press.

Visuoperceptual: basic visuoperceptual skills were assessed with the silhouette subtest of the Visual Object and Space Perception Battery (VOSP) (Warrington and James 1991).

Neuropsychological tasks were administered at each visit by me, in a single session lasting about 1½ hours. The order of tasks was kept constant. Alternative versions of the digit span and RMT were used at the six-month assessment in order to try and minimise practice effects.

3.2.3 Statistical Analysis

Group differences in age and estimated premorbid IQ at baseline, and scan interval, were investigated using t-tests assuming unequal variance. Fisher's

exact test was used to see if the proportion of right-handed subjects, or males, differed between groups. t-tests (assuming unequal variance where necessary) were also used to investigate whether the stage 1 and stage 2 HD subjects differed in terms of age or CAG repeat length.

In all the following analyses the effects of variables which can independently affect brain volume (age) and cognitive performance (age and estimated premorbid IQ) were controlled for by including them as covariates where necessary. The NART, which was used to estimate IQ, may be affected by disease (Solomon *et al.* 2007). If this is the case then including it as a covariate will tend to reduce the significance of group effects; however failing to include it would mean that group effects might be attributable to (non-disease-related) differences in IQ alone.

3.2.3.1 Clinical measures

Linear regression models were used to investigate the relationship between CAG repeat length and baseline UHDRS scores. Paired t-tests (two-tailed) were used to look for change in UHDRS scores over 12 months. Linear regression models were used to investigate the relationship between CAG repeat length and change in UHDRS scores.

3.2.3.2 Brain volume

To correct for differences in head size baseline brain volumes were standardised for total intracranial volume (TIV) which was measured according to the protocol described in Appendix 2. In controls log-transformed brain volumes were regressed on log-transformed TIVs to derive the slope of the relationship

between TIV and brain volume. This coefficient, β, was then used to adjust all baseline brain volumes for TIV using equation 3.1.

$$V_{standardised} = V_{raw}\left(\frac{meanTIV}{subject-specificTIV}\right)^{\beta} \quad (3.1)$$

Where brain volume is mentioned in the results it always refers to volume standardised to mean control TIV.

Linear regression models, with robust standard errors to allow variance to differ between groups, were used to compare whole-brain volume (standardised for TIV) between groups at baseline, controlling for age by including it as a covariate. Linear regression models were also used to see whether there was a significant relationship between brain volume and CAG repeat length, and baseline UHDRS motor, independence and total functional capacity scores (again adjusting for age).

In addition, scan pairs were registered and the BBSI was derived for each pair. Atrophy rates were analysed on a logarithmic scale according to equation 3.2 in order to ensure that doublings and halvings be treated as effects of equal magnitude; follow-up volume was calculated by subtracting the BBSI from baseline volume.

$$\left[\frac{\left(\ln\left(V_{follow-up}/V_{baseline}\right)\right)}{time}\right] \quad (3.2)$$

Regression models relating the log-transformed variables to subject group and age were used to obtain individual age-standardized atrophy rates. Geometric

mean atrophy rates (% per year) were calculated by back transformation with standard deviations (SDs) calculated from variance transformation formulae.

A linear regression model with robust standard errors was used to compare rates of whole-brain atrophy between groups. Although the effect of age on atrophy rates is thought to be relatively small in this age range (Scahill *et al.* 2003), age was included as a covariate in this model in order to ensure that any group differences seen could not be attributed to this. Atrophy rate was then regressed on CAG repeat length and change in UHDRS motor score, independence score and total functional capacity (again adjusting for age).

3.2.3.2.1 Comparison of the BBSI and manual measures

In addition to being estimated from the BBSI volume loss was also estimated from the raw volumes, and volumes standardised to mean control TIV. Also, because the control sample from which the relationship between TIV and volume was estimated was small and thus possibly not representative, volume loss was also estimated from volumes as a proportion of TIV. For each of these measures t-tests assuming unequal variance were used to assess the differences in the volume loss between groups.

The mean and standard deviation of loss estimated from the BBSI were compared with those estimated from raw volumes, and volumes standardised to mean TIV, using paired t-tests (means) and Pitman's tests (standard deviations). This was not done for the BBSI and volume as a proportion of TIV as these measurements have different units. Paired t-tests were also used to investigate whether TIV differed between timepoints.

3.2.3.3 Cognition

Linear regression models with robust standard errors were used to compare cognitive measures between groups at baseline, controlling for age and estimated premorbid IQ. For naming latency measures, the natural log of time to correct response was related to group using a linear mixed model with group, question, age and IQ as fixed effects and subject as a random effect. Provided that data are "missing at random" (Rubin 1976; Feudjo-Tepie *et al.* 2006) such an approach is unbiased even if questions that take longer on average are also more likely not to be answered. Linear regression models were also used to investigate the relationship between baseline cognitive performance and CAG repeat length, adjusting for age and estimated premorbid IQ.

For cognitive variables a change score was generated from baseline and follow-up scores, such that a negative score indicated decline over time. Linear regression models with robust standard errors were used to compare change scores between groups, controlling for baseline age and estimated premorbid IQ.

Associations between change in cognitive variables and CAG repeat length were assessed using linear regression models, adjusting for age and estimated premorbid IQ. Within the HD group associations between change in cognitive variables and atrophy rate were also assessed using linear regression models.

3.3 RESULTS

Early HD subjects and controls were well-matched for demographic variables at baseline, including scan interval. There were small non-significant differences in age and IQ, gender and handedness (Table 3-1). Age did not differ significantly between the stage 1 and stage 2 HD subjects (mean (SD) age 46.0 (7.1) years vs.

45.0 (4.4) years, p=0.78). Stage 2 subjects tended to have higher repeat lengths than stage 1 but this difference was not statistically significant (stage 1: mean (SD) 44.6 (1.6) repeats vs. stage 2: 47.0 (2.0) repeats, p=0.06).

Table 3-1 Demographic data for the cohort at baseline

	Control (N=7)	Early HD (N=13)
Gender (M:F)	3:4	5:8
Age (year)	40.7 (10.5)	45.8 (6.4)
Estimated premorbid IQ	106.6 (10.2)	100.4 (11.9)
Handedness (R:L)	7:0	11:2
CAG repeat length	NA	45.2 (2.0), range 43 – 49
Disease duration (year)	NA	3.5 (2.7), range 1 – 8
UHDRS Motor	NA	12.7 (12.7)
UHDRS Independence	NA	93.5 (10.5)
UHDRS TFC	NA	11.6 (2.6)
Scan interval (months)	5.8 (0.6)	6.0 (0.8)

NA: not applicable

Data are mean (SD) with the exception of gender and handedness; handedness was taken as the hand used to write with; UHDRS: motor is out of 124, higher score = more severely impaired; independence is scored as a percentage, higher score = better function; Total Functional Capacity is out of 13, higher score = better function.

3.3.1 Clinical measures

At baseline longer CAG repeat length was associated with worse scores on the UHDRS motor task (4.6 points per CAG repeat, 95% CI 1.6, 7.7, p=0.007), independence scale (-3.7 points per CAG repeat, 95% CI -6.3, -1.0, p=0.011) and TFC (-0.9 points per CAG repeat, 95% CI -1.5, -0.2, p=0.017).

At 12-month follow-up one early HD subject had progressed from stage 2 to stage 3. For the entire patient group there was a statistically significant increase in UHDRS motor score (p=0.0001) over 12 months, indicating a decline in motor skills. Independence score and TFC both decreased over 12 months (Independence score, p=0.0145, TFC, p<0.0075) (Table 3-2).

Table 3-2 Mean (SD) UHDRS scores in the early HD group at baseline and 12-month assessments, with change (95% confidence intervals)

	Early HD (N=13)		Change
	Baseline	12 months	(12 months - baseline)
Motor score	12.7 (12.7)	15.9 (14.0)	3.2 (2.0, 4.5) p=0.0001
Independence score	93.5 (10.5)	89.2 (14.4)	-4.2 (-7.5, -1.0) p=0.0145
TFC	11.6 (2.6)	10.7 (3.4)	-0.9 (-2.6, -0.3) p=0.0075

Longer CAG repeat length was associated with greater decline at motor score (decline increased by 0.6 points per CAG repeat, 95% CI 0.1, 1.2, p=0.032) and TFC (decline increased by 0.3 points per CAG repeat, 95% CI 0.02, 0.6, p=0.037), but there was no evidence of an association between CAG repeat length and decline in independence score.

3.3.2 Brain volumes

HD subjects had smaller mean brain volumes than controls (standardised to mean TIV) at baseline (1040 vs. 1132 ml), both with and without adjustment for age (both p<0.0001, Table 3-3). There was no evidence of a relationship between brain volume and CAG repeat length, UHDRS motor score, independence score or TFC. Raw volumes and TIVs are reported in the following section, 3.3.2.1.

Table 3-3 Mean (SD) baseline brain volumes, standardised to mean control TIV, and rate of atrophy over six months, with differences (95% confidence intervals) with and without adjustment for age

	Control	Early HD	Difference (Early HD - Control)	
	(N=7)	(N=13)	Crude	Adjusted
Brain volume (ml)	1132 (40)	1040 (46)	-92 (-133, -51) p<0.0001	-83 (-118, -47) p<0.0001
Atrophy rate (% per year)[a]	0.26 (0.54)	1.10 (0.88)	0.84 (0.17, 1.50) p=0.017	0.87 (0.19, 1.55) p=0.015

Negative differences indicate smaller volume or atrophy rates in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age; adjusted differences are the differences having adjusted for age by including it as a covariate in the linear regression models.

a Loss over six months is presented as a percentage of baseline brain volume lost per year

HD subjects also had greater mean rates of atrophy than controls, both with and without adjustment for age (without age p=0.017; with age p=0.015) (Table 3-3, Figure 3-1). There was no evidence of a relationship between atrophy rate and CAG repeat length, or decline in UHDRS motor score, independence score or TFC.

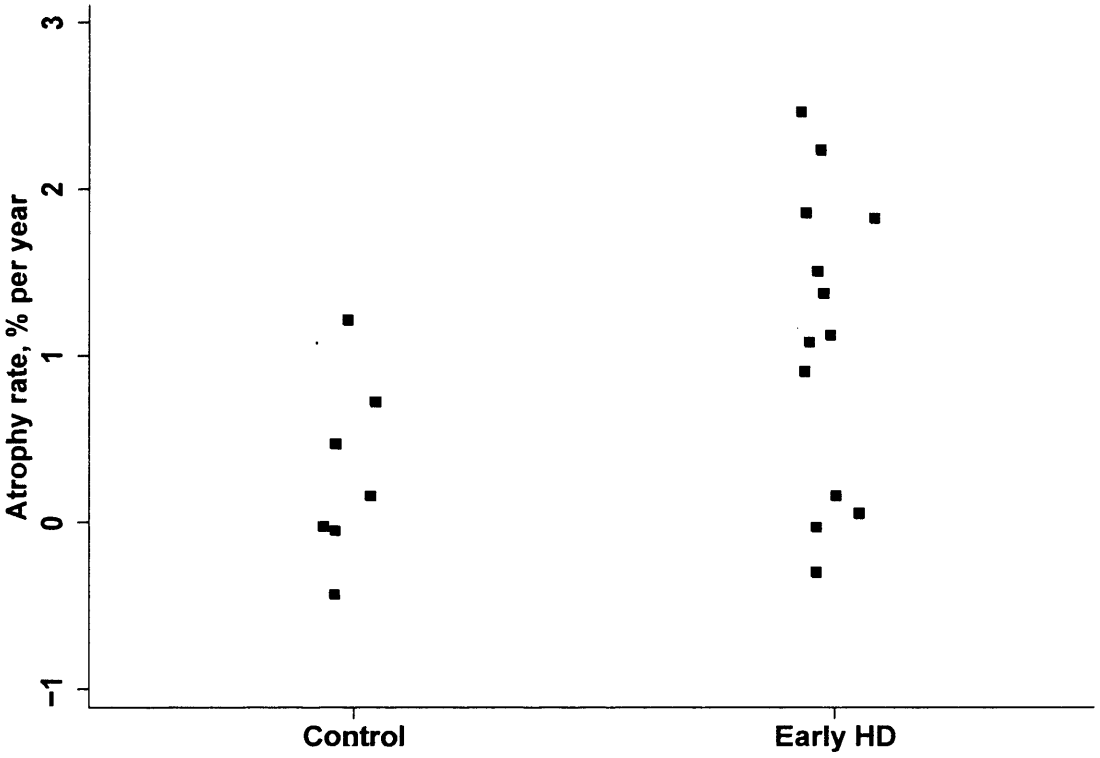


Figure 3-1 Whole-brain atrophy rates (% per year) in controls and early HD, standardised to mean age in the whole cohort

3.3.2.1 Comparison of BBSI and manual measures

All four measures of volume loss (from raw volumes, volumes standardized to mean control TIV, volumes as a proportion of TIV, and the BBSI) showed loss over six months, and all demonstrated statistically significant differences in the amount of change between controls and HD (see Table 3-4).

Table 3-4 Mean (SD) brain and total intracranial volumes and loss over 6 months from manual measures and the BBSI, with group differences (95% confidence intervals)

		Baseline		6 months		Loss over 6 months		Group difference in loss
		Control	Early HD	Control	Early HD	Control	Early HD	Early HD - Control
TIV (ml)		1370 (81)	1393 (144)	1359 (82)	1377 (135)	-10.8 (11.3)	-15.2 (14.6)	-4.44 (-16.98, 8.09) p=0.46
Brain volume (ml)	Raw[a]	1124 (59)	1043 (94)	1119 (60)	1023 (90)	-4.87 (6.98)	-20.29 (8.05)	-15.42 (-22.83, -8.01) p=0.0005
	Log[b]	1132 (40)	1040 (46)	1126 (40)	1021 (46)	-6.33 (2.63)	-19.67 (6.74)	-13.34 (-17.80, -8.87) p<0.0001
	Prop[c]	0.82 (0.03)	0.75 (0.04)	0.82 (0.03)	0.74 (0.04)	0.003 (0.003)	-0.007 (0.006)	-0.010 (-0.014, -0.005) p=0.0002
BBSI (ml)		-	-	-	-	-1.47 (2.98)	-6.16 (5.02)	-4.68 (-8.45, -0.91) p=0.0178

Negative loss indicates smaller volume at six-month follow up than at baseline; negative group differences indicate that the HD group lost more than the control group.

a Raw volumes are uncorrected

b Log volumes are standardised for head size assuming a linear relationship between log transformed values and log TIV, see text for details

c Proportional volumes are represented as a proportion of TIV (volume/TIV); unlike log correction, this method does not rely on the provision of normative data from the control group

Mean volume loss estimated by the BBSI was smaller than that calculated from differences between raw or TIV-corrected volumes (BBSI vs. raw, 10.4ml, 95% CI 6.7, 14.1ml, p<0.0001; BBSI vs. volumes standardised to mean control TIV, 10.5ml, 95% CI 6.8, 14.2ml, p<0.0001). The standard deviation of loss estimated by the BBSI was also smaller than that calculated from the other two methods (BBSI vs. raw, ratio of SDs=0.5, 95% CI 0.3, 0.6, p<0.0001; BBSI vs. volumes standardised to mean control TIV, ratio of SDs=0.6, 95% CI 0.4, 0.9, p=0.013).

In the subject group as a whole TIVs were also smaller at six months compared to baseline (-13.7ml, 95% CI 7.4, 20.0ml, p=0.0002). However there was no evidence that the change in TIV over six months differed between early HD subjects and controls (Table 3-4).

3.3.3 Cognitive performance

Differences in baseline neuropsychology, with and without adjustment for age and estimated premorbid IQ, are shown in Table 3-5. After adjustment HD subjects were generally worse than controls on all tasks, with the exception of digit span and naming in the GNT where they tended to perform slightly better than controls. Early HD subjects were statistically significantly worse than controls at Stroop interference (p=0.017), phonemic fluency (p=0.029) and the SDMT (p=0.038) as well as recognition memory for faces (RMF) (p=0.009). They were significantly slower at number cancellation (p=0.015), number sequencing (p=0.007), letter sequencing (p=0.034), alternating sequencing (p=0.048), motor speed (p=0.011) and the reaction time task (p=0.005). Although HD subjects' naming performance was unimpaired they had

significantly slower mean naming latency than controls in the GNT ($p<0.0001$). Differences between HD and controls on Stroop word reading, digit span forwards and backwards, recognition memory for words (RMW), the recognition memory difference score and the silhouette subtest of the VOSP were not statistically significant (all $p>0.1$).

Longer CAG repeat length was generally associated with worse or slower baseline scores (after adjusting for age and premorbid IQ), and was significantly associated with slower times on number sequencing (4.9 sec per CAG repeat, 95% CI 0.0, 9.8 sec, $p=0.049$) and reaction time (53.3 msec per CAG repeat, 95% CI 1.5, 105.0 msec, $p=0.045$). There was a trend towards a statistically significant relationship with worse performance on the SDMT (-4.0 points per CAG repeat, 95% CI -8.0, 0.1, $p=0.055$) and RMF (-1.0 points per CAG repeat, 95% CI -2.1, 0.1, $p=0.06$), and no evidence of relationships with the other cognitive tasks (all $p>0.09$).

Table 3-5 Mean (SD) cognitive scores at baseline, with differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ

		Controls	Early HD	Difference (Early HD - Control)	
Area	Test	(N=7)	(N=13)	Crude	Adjusted
UHDRS	Stroop word (n)	91.4 (20.8)	74.5 (23.9)	-17.0 (-38.3, 4.4) p=0.11	-7.7 (-34.0, 18.7) p=0.55
	Stroop interference[a] (n)	49.0 (6.6)	34.7 (10.7)	-14.3 (-22.7, -5.9) p=0.002	-9.3 (-16.7, -1.9) p=0.017
	Phonemic fluency (n)	50.6 (16.3)	33.5 (13.9)	-17.0 (-32.1, -2.0) p=0.029	-16.1 (-30.4, -1.8) p=0.029
	SDMT (n)	55.7 (14.7)	35.8 (12.5)	-19.9 (-33.5, -6.4) p=0.006	-15.0 (-28.9, -1.0) p=0.038
Psychomotor,	Number cancellation (sec)[b]	17.9 (3.3)	25.2 (5.7)	-7.3 (-11.5, -3.0) p=0.002	-5.3 (-9.4, -1.2) p=0.015
executive	Number sequencing (sec)[b]	24.3 (8.1)	45.7 (16.3)	-21.5 (-32.9, -10.0.) p=0.001	-16.3 (-27.5, -5.0) p=0.007
function,	Letter sequencing (sec)[b]	25.9 (7.7)	50.3 (28.6)	-24.4 (-42.3, -6.5) p=0.01	-14.1 (-27.0, -1.2) p=0.034
reaction time	Alternating (sec)[b]	69.4 (34.1)	169.4 (132.9)	-100.0 (-182.8, 17.3) p=0.069	-69.5 (-138.5, -0.6) p=0.048
	Motor speed (sec)[b]	23.9 (5.6)	34.7 (10.5)	-10.8 (-18.4, -3.3) p=0.008	-7.3 (-12.6, -1.9) p=0.011
	Reaction time (msec)	590 (104)	808 (158)	-218 (-342, -94) p=0.002	-187 (-307, -67) p=0.005

Area	Test	Controls (N=7)	Early HD (N=13)	Difference (Early HD - Control) Crude	Adjusted
Memory	Digit span forwards (/21)	11.7 (3.1)	13.2 (3.1)	1.4 (-1.6, 4.5) p=0.34	2.3 (-1.0, 5.5) p=0.16
	Digit span backwards (/21)	9.6 (2.1)	10.3 (4.9)	0.7 (-2.5, 4.0) p=0.64	2.3 (-1.2, 5.9) p=0.19
	RMF (/50)	43.7 (4.0)	37.1 (4.4)	-6.6 (-10.7, -2.6) p=0.003	-6.3 (-10.8, -1.8) p=0.009
	RMW (/50)	47.9 (2.1)	43.8 (4.9)	-4.1 (-7.4, -0.8) p=0.019	-2.5 (-5.6, 0.6) p=0.11
	RMT difference	4.1 (5.5)	6.7 (3.0)	-2.5 (-7.1, 2.0) p=0.26	-3.8 (-8.8, 1.2) p=0.13
Language	GNT (/30)	20.3 (6.1)	20.7 (4.7)	0.4 (-4.7, 5.5) p=0.87	2.2 (-2.9, 7.3) p=0.37
	Naming latency (ln(msec))	7.5 (0.5)	7.9 (0.7)	-0.5 (-0.6, -0.3) p<0.0001	-0.3 (-0.5, -0.2) p<0.0001
Visuoperceptual	VOSP silhouettes (/15)	10.1 (2.2)	9.1 (2.3)	-1.1 (-3.3, 1.1) p=0.32	-0.6 (-2.4, 1.3) p=0.54

Negative differences indicate a lower score or slower speed (longer time taken in timed tasks) in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models.

a N=14, colour-blind subject unable to complete test

b Subtests of the Delis-Kaplan Executive Function System (D-KEFS) Trail Making Test

Between baseline and six months, HD subjects tended to decline relative to controls at all tasks (see Table 3-6), except Stroop colour-word, phonemic fluency and naming latency at which, on average, they improved. HD subjects only declined statistically significantly relative to controls at motor speed (p=0.005). Both groups had faster naming latencies at six months, with HD subjects showing significant improvement relative to controls (p=0.002), probably because controls were performing nearer to ceiling at baseline and hence had less scope to improve. This pattern was repeated between baseline and 12 months with HD subjects showing, on average, non-significant decline relative to controls on all tasks other than Stroop word reading, number sequencing, reaction time and naming latency (Table 3-7). HD subjects again declined statistically significantly relative to controls at motor speed (p=0.028) as well as the RMF (p=0.02) and RWM (p=0.004). Again although both groups named objects faster than at baseline, the HD group showed a greater improvement than controls (p=0.001). There were no other statistically significant changes between visits.

Longer CAG repeat length was generally associated with greater decline in cognition over both six and 12 months (after adjusting for age and estimated premorbid IQ). Over six months higher CAG repeat length was statistically significantly associated with greater decline at Stroop colour (decline increased by 5.4 points per CAG repeat, 95% CI 2.4, 8.4, p=0.003), cancellation (decline increased by 2.2 sec per CAG repeat, 95% CI 1.0, 3.5, p=0.002), and over twelve months with greater decline at Stroop colour (decline increased by 3.5 points per CAG repeat, 95% CI 1.2, 5.8, p=0.007). There was no evidence of associations between CAG repeat length and decline in other cognitive variables.

Table 3-6 Mean (SD) cognitive change from baseline to six months, with group differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ at baseline

		Change (6 months - baseline)		Difference in change (Early HD - Control)	
Area	Test	Controls (N=7)	Early HD (N=13)	Crude	Adjusted
UHDRS	Stroop word (n)	3.3 (14.7)	-0.8 (12.1)	-4.1 (-17.6, 9.3) p=0.53	-7.4 (-22.0, 7.2) p=0.30
	Stroop interference[a] (n)	-0.9 (3.5)	0.3 (5.0)	1.1 (-3.0, 5.3) p=0.58	1.3 (-3.2, 5.8) p=0.55
	Phonemic fluency (n)	-1.7 (6.6)	1.2 (5.4)	2.9 (-3.1, 8.9) p=0.33	4.0 (-2.9, 10.8) p=0.24
	SDMT (n)	0.6 (6.7)	-1.4 (5.1)	-2.0 (-8.0, 4.1) p=0.50	-0.7 (-6.0, 4.5) p=0.77
Psychomotor,	Number cancellation (sec)[b]	1.8 (3.6)	-1.2 (4.9)	-3.0 (-7.0, 1.0) p=0.14	-3.9 (-8.5, 0.7) p=0.09
reaction time	Number sequencing (sec)[b]	0.7 (6.1)	2.9 (16.1)	2.2 (-8.5, 12.8) p=0.68	-1.0 (-12.2, 10.2) p=0.86
	Letter sequencing (sec)[b]	4.2 (4.3)	-4.2 (30.8)	-8.4 (-26.9, 10.1) p=0.35	-8.1 (-22.0, 5.9) p=0.24
	Alternating (sec)[b]	-9.4 (35.6)	-0.4 (74.3)	9.0 (-42.8, 60.8) p=0.72	-3.6 (-83.8, 76.7) p=0.93
	Motor speed (sec)[b]	5.6 (3.2)	-1.5 (7.8)	-7.1 (-12.3, -1.9) p=0.011	-7.5 (-12.4, -2.6) p=0.005

Area	Test	Change (6 months - baseline)		Difference in change (Early HD - Control)	
		Controls (N=7)	Early HD (N=13)	Crude	Adjusted
	Reaction time (msec)	-1 (25)	-20 (64)	-19 (-61, 24) p=0.37	-16 (-68, 37) p=0.54
Memory	Digit span forwards (n)	1.1 (1.3)	-1.2 (3.4)	-2.4 (-4.6, -0.1) p=0.04	-1.9 (-4.1, 0.4) p=0.10
	Digit span backwards (n)	0.4 (3.5)	-1.6 (2.3)	-2.0 (-5.0, 1.0) p=0.17	-1.8 (-6.2, 2.6) p=0.39
	RMF (n)	1.6 (4.5)	-2.3 (4.8)	-3.9 (-8.3, 0.6) p=0.08	-2.2 (-7.6, 3.3) p=0.41
	RMW (n)	0.4 (2.4)	-0.4 (3.6)	-0.8 (-3.7, 2.0) p=0.56	-1.3 (-5.0, 2.3) p=0.45
Language	GNT (n)	1.6 (1.4)	0.5 (1.4)	-1.0 (-2.4, 0.3) p=0.13	-1.0 (-2.5, 0.4) p=0.16
	Naming latency (ln(sec))	5.5 (1.6)	6.6 (1.6)	1.0 (0.6, 1.5) p<0.0001	0.8 (0.3, 1.3) p=0.002
Visuoperceptual	VOSP silhouettes (n)	0.9 (1.5)	0.5 (1.9)	-0.3 (-1.9, 1.3) p=0.68	-0.5 (-2.7, 1.6) p=0.60

Negative change indicates a decrease in score or in speed (longer time taken in timed tasks); negative difference in change indicates that early HD declined more than controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models.

a N=14, colour-blind subject unable to complete test

b Subtests of the Delis-Kaplan Executive Function System (D-KEFS) Trail Making Test

Table 3-7 Mean (SD) cognitive change from baseline to 12 months, with group differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ at baseline

		Change (12 months - baseline)		Difference in change (Early HD - Control)	
Area	Test	Controls (N=7)	Early HD (N=13)	Crude	Adjusted
UHDRS	Stroop word (n)	-2.4 (10.7)	-3.0 (8.3)	-0.6 (-10.2, 9.1) p=0.90	1.6 (-6.8, 10.0) p=0.69
	Stroop interference[a] (n)	0.7 (3.8)	-0.2 (3.4)	-0.9 (-4.5, 2.7) p=0.62	-1.4 (-4.8, 2.0) p=0.39
	Phonemic fluency (n)	2.0 (8.6)	-2.2 (5.5)	-4.2 (-11.7, 3.2) p=0.25	-2.8 (-13.3, 7.6) p=0.58
	SDMT (n)	0.9 (6.5)	-2.0 (5.0)	-2.9 (-8.7, 3.0) p=0.32	-1.7 (-9.7, 6.3) p=0.66
Psychomotor,	Number cancellation (sec)[b]	1.3 (5.1)	-0.4 (4.8)	-1.7 (-6.6, 3.2) p=0.47	-1.4 (-7.1, 4.3) p=0.60
reaction time	Number sequencing (sec)[b]	4.3 (6.8)	2.8 (16.7)	-1.5 (-12.7, 9.7) p=0.78	2.4 (-11.3, 16.2) p=0.71
	Letter sequencing (sec)[b]	4.0 (7.6)	8.3 (18.9)	4.3 (-8.3, 17.0) p=0.48	-1.3 (-11.7, 9.2) p=0.80
	Alternating (sec)[b]	4.0 (20.8)	6.5 (75.0)	2.4 (-44.7, 49.6) p=0.92	-5.2 (-61.5, 51.0) p=0.85
	Motor speed (sec)[b]	8.2 (4.0)	0.6 (9.4)	-7.6 (-13.9, -1.3) p=0.021	-7.5 (-14.1, -0.9) p=0.028

Area	Test	Change (12 months - baseline)		Difference in change (Early HD - Control)	
		Controls (N=7)	Early HD (N=13)	Crude	Adjusted
	Reaction time (msec)	14 (68)	-0.8 (146)	-15 (-116, 86) p=0.76	8 (-108, 124) p=0.89
Memory	Digit span forwards (n)	-0.3 (2.8)	-1.4 (3.4)	-1.1 (-4.1, 1.9) p=0.45	-0.8 (-3.6, 2.1) p=0.57
	Digit span backwards (n)	2.3 (2.9)	-0.5 (2.6)	-2.8 (-5.6, -0.1) p=0.044	-2.9 (-6.8, 0.9) p=0.12
	RMF (n)	2.3 (2.5)	-2.2 (5.0)	-4.5 (-8.1, -1.0) p=0.015	-3.5 (-6.4, -0.6) p=0.02
	RMW (n)	0.6 (1.9)	-0.8 (2.0)	-1.3 (-3.2, 0.6) p=0.15	-2.5 (-4.1, -0.9) p=0.004
Language	GNT (n)	1.3 (1.3)	0.8 (1.3)	-0.4 (-1.7, 0.9) p=0.49	-0.3 (-1.8, 1.1) p=0.65
	Naming latency (ln(sec))	5.8 (1.5)	6.6 (1.5)	0.9 (0.5, 1.3) p<0.0001	0.8 (0.3, 1.3) p=0.001
Visuoperceptual	VOSP Silhouettes (n)	1.0 (1.4)	0.7 (1.7)	-0.3 (-1.8, 1.2) p=0.67	-0.2 (-1.9, 1.5) p=0.81

Negative change indicates a decrease in score or in speed (longer time taken in timed tasks); negative difference in change indicates that early HD declined more than controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models.

a N=14, colour-blind subject unable to complete test

b Subtests of the Delis-Kaplan Executive Function System (D-KEFS) Trail Making Test

In general atrophy rate was not significantly associated with decline in cognitive variables over six or 12 months (after adjusting for age and estimated premorbid IQ). Higher atrophy rate was associated with less decline in motor speed over six months (5.3 sec per % increase in atrophy rate, 95% CI 0.8, 9.7, p=0.025) and 12 months (6.1 sec per % increase in atrophy rate, 95% CI 0.6, 11.6, p=0.033) and with less decline at GNT over six months (0.8 points per % increase in atrophy rate, 95% CI 0.02, 1.5, p=0.046).

3.4 DISCUSSION

In this sample of subjects with early HD, whole-brain atrophy occurred at a mean rate of around 1% per year (based on the finding of 0.5% loss over 6 months) and the difference between this and atrophy due to normal ageing can be detected over this period, even with the very small groups in this study. Despite the group differences there was considerable overlap between groups. This could reflect true heterogeneity or may be in some part due to factors that were not controlled for (such as hydration changes) or errors related to the subject (movement artefact), scanner or measurement technique. The rate in early HD was around four times greater than that in controls. These rates of loss together with the baseline cross-sectional volume differences suggest that atrophy has been occurring for some years prior to first assessment.

Whole-brain atrophy rates are slower than striatal rates in early HD (around 5% per year in the caudate) (Aylward *et al.* 2000), consistent with the assumption that whole-brain volume includes areas that are not changing, together with those that are. Whole brain losses of 1% per year (about 10ml per year) cannot be attributed simply to striatal atrophy; total caudate loss is approximately 0.5ml per

year based on published atrophy rates (Aylward *et al.* 2000). At post-mortem in end-stage HD there is a loss of 300-400 grams of brain tissue, or between 20-30% of brain volume in HD (de la Monte *et al.* 1988), most of which must represent extra-striatal atrophy.

Cognitive deficits at baseline were similar to those reported by others (e.g., Craufurd and Snowden 2002), with the HD group tending to perform below control levels, significantly so on a number of motor and executive tasks, and recognition memory for faces. In contrast naming and visuoperceptual skills appeared relatively spared. Interestingly, although there was no evidence of a naming deficit in the HD group, early HD subjects had significantly longer naming latencies than controls, most likely indicative of general cognitive slowing, and also highlighting the power of timed tests.

Performance on UHDRS clinical measures declined significantly over the year, whilst early HD subjects showed slowed motor speed and tended to get worse at recognition memory for faces and words, relative to controls. However these changes were not significantly associated with atrophy rates. In addition, it is clear that measurement of cognitive change is variable, particularly over short time periods, since decline over six months was slightly different to that seen over 12 months, and effects were generally small and non-significant; these data must therefore be interpreted cautiously and may not be representative of the disease. A number of other studies have reported difficulties in detecting cognitive change over even longer time periods than the ones used here, primarily because of practice effects and performance variability (e.g., Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001). The importance of testing change against

control performance (rather than against zero) is demonstrated, as controls showed practice effects in some tests.

Higher CAG repeat length was generally associated with worse clinical and cognitive scores at baseline, and with greater decline in some of the UHDRS measures, although only with decline in one cognitive measure (Stroop colour reading). CAG repeat length was not significantly associated with brain volume, atrophy rate or most cognitive change, which may be indicative of the fact that CAG repeat length does not govern rate of disease progression; alternatively this may simply reflect the small sample size and lack of power to detect small effects.

In terms of the practical aims of the study, the results show that the BBSI can be used to detect change in early HD, and that serial MRI is well tolerated, at least in the cohort tested here. There was no loss to follow-up at any of the three timepoints suggesting that subjects were happy to return for repeat assessments despite progression of the disease in the intervening months for the early HD subjects. However, HD is a phenotypically heterogeneous disease (Shoulson and Fahn 1979), which makes it hard to find consistent patterns of deficit. The whole-brain atrophy rates reported here reflect that, in that they were highly variable and included a number of patients well within the control range. This was mirrored by the cognitive performance, in which decline was variable and hard to detect, and what significant change there was, was often attributable to a practice effect in the controls as well as a small decline in the HD group.

These difficulties were compounded by the fact that the scanner upgrade forced a change in the study schedule, and meant that atrophy rates were measured over

six months (rather than one year) and that subjects underwent the cognitive assessment more frequently than was ideal, making it hard to minimise practice effects. The short time period meant a high signal-to-noise ratio, which is likely to have added to the variance associated with the cognitive and imaging measures. As this was a pilot study the sample size was also small, and again this means the study was unlikely to have the power to detect small effects.

Four scan pairs (20%) were not adequate due to artefact and were discarded from the analysis. It is common to assume 10% loss to follow-up and 10% unusable scans in large-scale trials (Fox *et al.* 2000), so in this study slightly more scan pairs were unusable than would normally be predicted, although because no subjects were lost to follow-up the overall drop-out rate was similar to that seen in large-scale trials.

It is known that the BBSI tends to underestimate loss slightly (Freeborough and Fox 1997), and in this cohort the mean loss estimated by the BBSI was significantly smaller than that estimated by differences between volumes (both raw, and standardised to mean control TIV). However mean TIV also decreased significantly between timepoints, which suggests that some scanner drift had occurred. TIV, an index of head size, would not normally be expected to decline over time. It seems likely that scanner drift contributed partly to the large differences between brain volumes measured manually at each timepoint, and demonstrates the importance of registration in serial imaging, since this reduces such effects. This issue is discussed further in chapter 10, in terms of the "best" manual measure of volume change with which to compare the BBSI.

Whole-brain atrophy is simply and reliably derived from standard volumetric MRIs and the BBSI. This semi-automated measure is less susceptible to segmentor error than manual delineation, and it is not affected by changes in scanner calibration which are often seen with unregistered scan pairs; this is reflected in the reduced within-group variance and (in this cohort) slightly smaller change compared to manual measures. The findings of this study suggest that the BBSI may be a practical measure for use in larger-scale studies of HD, whilst the detection of cognitive change would be improved with a longer time period and more sensitive tests.

3.5 CONCLUSION

This study has shown that the BBSI can be used to measure whole-brain atrophy in early HD, although there are a number of challenging aspects to overcome, such as movement artefact and ensuring scanner stability. Cognitive change is much harder to measure, partly because the disease is so clinically heterogeneous, and partly because of factors which cannot be controlled for such as practice effects and subjects' mood. Because of the short time period and small cohort I have not attempted to interpret the cognitive findings as representative of the decline in HD in general. Rather, these findings lead on to the work reported in the following chapters in which a number of limiting factors were addressed: the sample size was increased; the time interval was lengthened to increase the signal-to-noise ratio and reduce cognitive practice effects; the cognitive battery was adapted; and premanifest subjects were recruited in order to expand these investigations across the wider spectrum of the disease.

4 COGNITIVE DEFICITS AND THEIR RELATIONSHIP WITH BRAIN VOLUME IN HUNTINGTON'S DISEASE

4.1 INTRODUCTION

As discussed in chapter 1, there is increasing evidence that subtle cognitive and behavioural deficits can be apparent prior to motor onset and that atrophy in HD is not just confined to the striatum, even in the early stages (Kassubek *et al.* 2004a; Williams *et al.* 2007). Defining these changes better allows one to a) have a greater understanding of "clinical onset" that captures the heterogeneity of the disease beyond motor deficits and b) thereby develop measures of progression that would be most appropriate for disease-modification trials aimed at the very earliest symptomatic stages of the disease.

Premanifest subjects are the focus of much research, in part because they are still "well" and are the most likely candidates for clinical trials, but this brings its own difficulties. Firstly, many of these subjects perform in the normal range on most clinical or cognitive measures, and thus it is hard to detect effects. Secondly, the term "premanifest" covers an extremely wide range of subjects, and until recently it has been difficult to predict with any accuracy, how far a premanifest subject was from motor onset (see section 1.4.3). This has made it hard to generalise results between studies of premanifest subjects, and to the population as a whole.

Following on from the pilot study (chapter 3) a larger study was set up, the baseline assessment of which was designed to address some of the above issues. This chapter examines the cognitive impairments seen in a large cohort of premanifest and early HD subjects, using an extensive battery of cognitive tasks. Whole-brain volume was measured to investigate the association between

cognitive performance and neurodegeneration. Also, since the classification of gene carriers into premanifest and early HD is made purely on the basis of motor signs, which do not necessarily correlate with cognitive ability, this chapter also investigates the association between cognition and brain volume across the whole spectrum of gene carriers. This study aimed to establish both the cognitive impairments and volume reductions in a large, well-defined cohort, and the associations between the two.

4.2 METHODS

4.2.1 Subjects

Sixty-one subjects with genetically-confirmed HD were recruited. Forty were classified as patients with "early HD", 26 stage 1 and 14 stage 2, (Shoulson and Fahn 1979), and 21 subjects were gene carriers without motor signs, i.e. premanifest (PM). Throughout the rest of this thesis the "HD" group refers to the early (symptomatic) HD group; the premanifest group is referred to as that. On the rare occasions that both groups are combined for analysis subjects are referred to as "all HD gene carriers".

Twenty neurologically normal controls, either patients' partners (N=15) or non-gene carriers from at-risk families (N=5), were also recruited. Seven early HD subjects and one control had previously participated in the study described in chapter 3.

4.2.2 Assessments

4.2.2.1 Clinical assessment

Subjects were clinically assessed as described in section 2.2. Four early HD subjects were not asked to complete the Beck Depression Inventory (BDI) as it was not available to the clinician at the time of their assessments.

4.2.2.2 Cognitive assessment

Subjects underwent detailed neuropsychological testing. The previous battery (see section 3.2.2.3) was modified based both on the pilot study findings and discussion with other psychologists working in HD.

The European HD network (EHDN, www.euro-hd.net) was set up to establish a database of HD subjects across Europe and to discuss current assessment methods with an aim to getting a Europe-wide consensus on the tests used and to unify administration and scoring, in order to have a useful program of data collection in place before the start of large-scale clinical trials. The Cognitive Working Group of the EHDN has held several meetings to discuss the standard minimal battery. The cognitive tests included in the UHDRS are the Stroop test, phonemic (letter) fluency and the Symbol Digit Modalities Test. Whilst these were felt to be useful, they are all timed psychomotor tasks and it was felt that they did not provide a broad enough measure of skills known to be affected in the disease. In addition there are no validated alternative versions, so with repetition these tasks become subject to practice effects. Discussion within the working group led to the recommendation that a number of extra tests should be included in the minimum battery; these tests and the reasons for their inclusion are listed below.

Neuropsychological tasks were administered by me at each visit, in a single session lasting about 1½ to 2 hours. The order of the tasks was kept constant. Because a follow-up assessment was planned at a later date alternative test versions were used, where possible, on half the subjects (counterbalanced across subjects within each group). Tests with alternative versions were Spot the Word, Hopkins Verbal Learning Test, digit span, Homophone Meaning Generation Test and RMT.

Estimated Pre-morbid IQ: the NART was used, as in the Pilot Study (chapter 3). However a second measure, the Spot the Word Test, was also included (Baddeley *et al.* 1993). This is a lexical decision task in which subjects are shown pairs of word / non-words and asked to mark the real word. As in the NART the real words are uncommon, and scores on this task correlate highly with IQ. It was thought that the Spot the Word Test could have an advantage over the NART as a measure in HD patients because some patients with dysarthria might be penalised for poor pronunciation in the NART.

Executive function / psychomotor tasks: the UHDRS cognitive tests were unchanged from the pilot study, comprising the Stroop test, phonemic (letter) fluency, and the SDMT. However the version of the Stroop test had recently been changed to include separate colour-naming and word-reading sections prior to the interference task. Category fluency (Benton and Hamsher 1978) was added to the battery at the suggestion of the EHDN working group. Many researchers had found performance on this to be less affected until later in the disease, and therefore it might detect decline at a point at which most subjects were performing at floor level on phonemic fluency.

Trail Making parts A and B (TMT A and B) (Reitan and Wolfson 2004) were also recommended; they are quick to administer and the more simple part (number sequencing) can be attempted by subjects even if they find the subsequent number-letter sequencing part too difficult. Consequently TMT A and B were substituted for the more complex five-part test that was used in the pilot study (Delis *et al.* 2001), since performance on only one of these subtests (motor speed) had declined significantly during the pilot.

A simple, timed letter-cancellation task was included as a basic test of visual search and motor skills. An A4 page with a grid of the letters A, B, C, D, E in a pseudo-random order was presented, and subjects were asked to put a line through all the "As" as quickly as possible, without missing any.

The computer reaction time task was modified slightly. This task was designed and presented on SuperLab Pro 2.0 and required subjects to use the index finger of their dominant hand to press one of four numbered buttons on a response pad. Number presses were alternated with pressing of a central blue button. The numbers 1-4 were centrally presented on the screen in a pseudo-random order, alternating with a blue square, in 20 blocks of 4, with a pause after the 10th block. The experiment is demonstrated in Figure 4-1 on the following page.

Subjects were requested to press the correct button as quickly as possible in response to the on-screen stimulus and reaction time was recorded. After a correct button press a blank screen displayed for 500ms before the next stimulus. In the event of an error the stimulus stayed on the screen until the correct button was pressed, in order to ensure that all subjects had to press the central blue

button in between the number stimuli. Error trials were not included in the RT analysis.

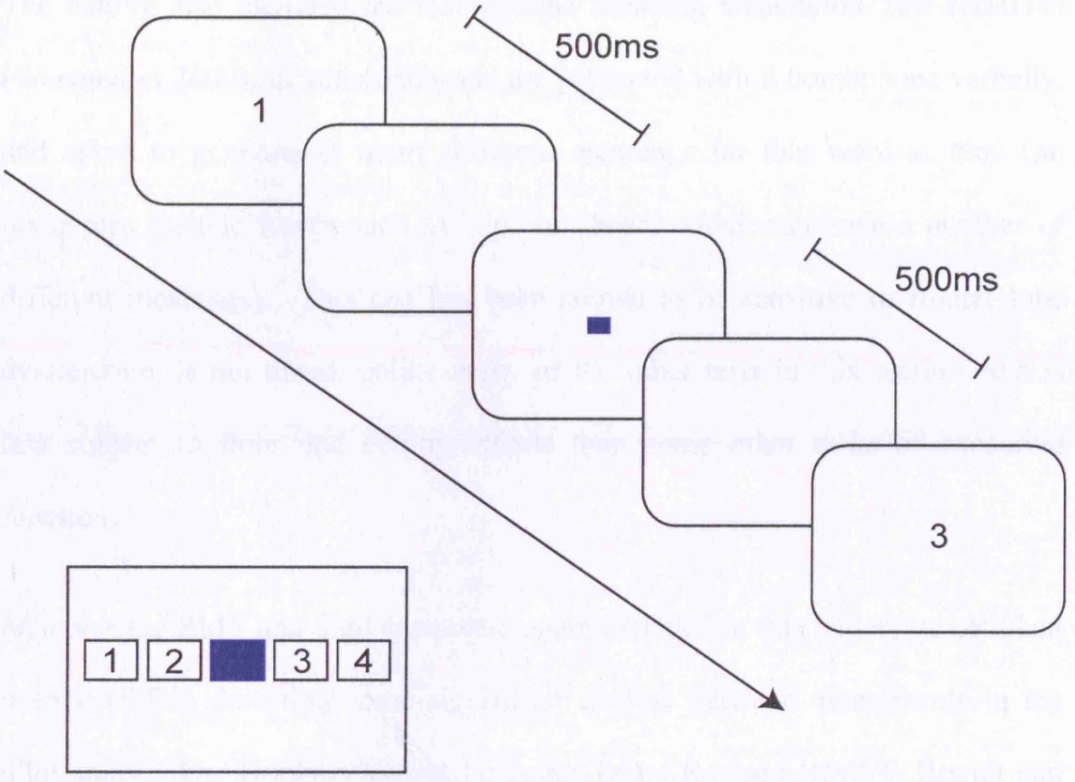


Figure 4-1 Diagram showing example screens from the RT experiment, and the layout of the button box

Subjects used the index finger of their dominant hand to press either a numbered button, or the central blue button, in response to the on-screen stimulus. The on-screen stimulus alternated between a number and a blue square with the numbers 1-4 presented randomly within each set of four number trials.

This modification of the experiment used in the pilot study was designed to yield four measures of RT: far (time to press buttons 1 and 4); near (time to press buttons 2 and 3); return from far (time to press central button after pressing 1 and 4); return from near (time to press central button after pressing 2 and 3). The use of the central blue button meant that subjects had to move from the same central

starting point to initiate the numbered button press and also allowed investigation of whether subjects with HD would be slow at returning to the central button, which was a predictable event.

The battery also included the Homophone Meaning Generation Test (HMGT) (Warrington 2000), in which subjects are presented with a homophone verbally, and asked to generate as many different meanings for that word as they can (examples include words such as "tip" or "bear" which can have a number of different meanings). This test has been shown to be sensitive to frontal lobe dysfunction, is not timed, unlike many of the other tests in this section, and is less subject to floor and ceiling effects than some other tasks of executive function.

Memory: the RMT and digit span were again included in this battery as both had been useful in detecting some significant decline between assessments in the pilot study. The Hopkins Verbal Learning Test - Revised (HVLT, Brandt and Benedict 2001) was included on the recommendation of the EHDN. Many researchers found this auditory verbal memory test useful, as it yields a number of measures (immediate memory, delayed memory, percentage recall, and a recognition discrimination index), and it is available in six alternative versions which helps to minimise practice effects.

Naming: the GNT was still included, and was again presented on a computer using SuperLab Pro 2.0 in order to measure latency of naming responses. In the pilot study latency measurements had been dependent on experimenter button press and in order to improve on the accuracy of this the experiment was modified slightly. Stimulus presentation was controlled by the experimenter, and

a sound was played at each stimulus onset. The whole test was recorded and the resulting .wav file was transferred to a Dell workstation. Goldwave software (www.goldwave.com) was used to insert markers between stimulus onset (the sound) and correct response, and naming latency was measured from this. Responses on which the subject was only correct after an error (e.g. mistaking the buoy for a sandcastle and then self-correcting) were not included in the latency analysis. Subjects were given unlimited exposure to stimuli although in practice subjects tended to indicate that they were unable to name an item within ~15 seconds of stimulus onset at which point the next stimulus was displayed.

Visuoperceptual: basic visuoperceptual skills were again assessed with the silhouette subtest of the VOSP.

Face processing and facial emotion recognition: the above tests comprise a comprehensive battery which was designed to cover the main areas of cognition in order to try and detect decline in HD. HD has also been used as a model for emotion recognition, although there are conflicting findings about which emotion is most impaired, and the extent of the impairment in premanifest gene carriers. Therefore emotion recognition was also investigated in this cohort. Facial emotion recognition was tested using 24 faces from the Ekman and Friesen battery (Ekman and Friesen 1976; after Gray *et al.* 1997). Subjects were shown a face and had to choose which of six basic emotions (written underneath) best described the face (see chapter 8 for more details). In order to control for face perception skills *per se*, the short form of the Benton Facial Recognition Test was also included in the battery (Benton *et al.* 1983). Total scores are reported in this chapter but a more detailed study of these data is presented in chapter 8.

4.2.2.3 MRI acquisition

Subjects underwent T1-weighted volumetric inversion recovery prepared, spoiled gradient recall scans on the same 1.5T GE scanner (for parameters see Appendix 1). Whole-brains and TIVs were segmented according to the protocol in Appendix 2.

Sixty-seven out of the 81 subjects (83%) had all three assessments (clinical, cognitive and MRI) on the same day. Eight early HD subjects had the clinical assessment after the cognitive and MRI due to the availability of the neurologist (mean (SD) interval 9.2 (6.0) weeks, range 1 – 22 weeks). Four controls were called back for rescans due to unacceptable motion artefact and hence their MR images were obtained after the other two assessments (mean (SD) interval 5.7 (2.6) weeks, range 2 – 7 weeks). One premanifest subject had the cognitive assessment three days after the clinical and MRI assessments. One premanifest subject had all three assessments on different days for logistical and rescan reasons, with the scan acquired five months after the cognitive data, and seven weeks after the clinical assessment.

4.2.3 Statistical Analysis

In the following analyses the effects of variables which could independently affect cognition (age and estimated premorbid IQ) and brain volume (age) were controlled for by including them as covariates. An exception to this was the Benton Facial Recognition score, which is already age- and education-adjusted.

4.2.3.1 Differences between alternative test versions

Linear regression models were used to compare performance between alternative versions of the psychological tests where they were available, across the cohort as a whole, adjusting for age and estimated premorbid IQ.

4.2.3.2 Group differences

Group differences in age, estimated premorbid IQ and CAG repeat length at baseline were investigated using t-tests assuming unequal variance. A chi^2 test was used to compare gender across groups. Fisher's exact test was used to examine whether the proportion of right-handed subjects differed between groups.

To correct for differences in head size whole-brains were standardised for TIV as outlined in section 3.2.3.2. Linear regression models with robust standard errors, to allow variances to differ between groups, were used to compare whole-brain volume (standardised for TIV), clinical and cognitive measures between groups at baseline. Linear regression models were also used for data which were not Normally distributed, but as Normality assumptions were violated 95% bias-corrected bootstrap confidence intervals with 1000 replicates were used; these data were from the RMW, TMT A and B, reaction time task, and the Ekman emotion recognition task.

Beck depression scores were used to classify subjects as showing signs of depression or not, and Fisher's exact test was used to see if the proportion of subjects classified as depressed differed between groups.

For naming latency measures, the natural log of time to correct response was related to group using a linear mixed model with group, question, age and IQ as fixed effects and subject as a random effect (see also section 3.2.3.3). This analysis was repeated with the UHDRS dysarthria score as an additional covariate to see if group differences were attenuated when the effect of dysarthria was adjusted for.

For TMT A and B, and the reaction time tasks, error counts were related to group using an overdispersed Poisson model to take into account the non-independence of errors made by the same person. A subset of errors on the reaction time task was classed as perseverative, if the subject repeated an incorrect button press, and these were modelled in the same way.

As there has been some debate as to the nature of the memory deficit in HD (see section 1.3.1.5) the HVLT was investigated in some detail. As above, linear regression models with robust standard errors were used to assess group differences in immediate recall, delayed recall, percent recalled, and recognition discrimination. In addition, similar models were used to investigate group differences in the numbers of true and false positives in the recognition paradigm, using 95% bias-corrected bootstrap confidence intervals with 1000 replicates as these scores were not Normally distributed. In order to test the theory that recognition performance is normal, given the number of items initially encoded, linear regression models were then used to assess group differences in delayed recall, discrimination, true and false positives, having adjusted for encoding by including the immediate recall score as a covariate. Unlike the percent recall score, which represents the proportion of encoded

stimuli which were recalled, this model tests the absolute difference between the number of stimuli encoded and recalled.

4.2.3.3 Effect of motor impairment on psychomotor tasks

In order to examine whether group differences in psychomotor tasks could be attributed to motor impairment, linear regression models with robust standard errors were used to look for differences between groups having controlled for UHDRS motor score by including it as a covariate.

4.2.3.4 Relationship between brain volume and clinical and cognitive measures in HD gene carriers

For the UHDRS independence score and TFC, at which the premanifest group scored at ceiling, linear regression models were used to assess the relationship between these measures and brain volume in the HD group alone, adjusting for differences due to age.

A linear regression model was also used to investigate the relationship between disease duration and brain volume in the early HD group, again adjusting for age. Disease duration is time since the diagnosis of motor signs (i.e., current age – age at onset). As duration and age linearly combine to give age at onset, this model implicitly investigates the effect of age at onset after adjustment for age as well as the effect of disease duration after adjustment for age. Current age needs to be adjusted for, for the reasons given above, and hence it is difficult to tease out the effects of duration and age at onset individually.

Linear regression models (with robust standard errors) relating either CAG repeat length, cognitive scores or UHDRS motor score, to brain volume, group and their interaction, were used to investigate whether the association between these variables and brain volume differed between the premanifest and HD groups. In situations where there were significant interactions the slope in each group was examined to see if, in either group, the association was significantly different from zero. Secondly, for variables for which there was no evidence of an interaction, linear regression models were used to look for an association between volume and score in all gene carriers, having adjusted for differences due to group. This ensures that group differences in score are not driving the correlation.

4.2.3.5 Relationship between probability of motor onset and other variables

Estimated probability of motor onset within five years in the premanifest group was calculated using the equation given by Langbehn *et al.* (2004). Probability of onset, rather than years to onset, was used as a regressor as it has been suggested that the former has more linear relationships with other clinical variables than the latter (Solomon *et al.* 2007). Where probability of onset is referred to in the thesis it always refers to estimated probability of onset within five years of baseline assessment. Linear regression models were used to see whether clinical, MRI or cognitive variables were associated with probability of motor onset in the 21 premanifest subjects. Age and IQ were included as covariates.

4.2.3.6 Comparison of NART and the Spot the Word Test

The Spot the Word age-scaled scores have a mean of 10 and standard deviation of three, and so they were transformed to be on the same scale as the NART IQ (mean of 100, standard deviation of 15). Agreement between the two measures was assessed by calculating the mean (SD) difference (Bland and Altman 1986). t-tests assuming unequal variance were used to assess whether the mean difference between the tests differed significantly between groups, and a paired t-test was used to see if the IQ estimates from each test were significantly different for the cohort as a whole. Pitman's test was used to compare variances of the two measures. The extent of association between the two measures was assessed using a correlation coefficient.

4.3 RESULTS

4.3.1 Group differences

There were small non-significant differences in gender, handedness and IQ between the groups (Table 4-1). Mean age did not differ statistically significantly between the controls and early HD group. The mean age of the premanifest group was seven and a half years lower than controls (95% CI 1.8, 13.5, p=0.011). The premanifest group had a shorter mean CAG repeat length than the early HD group (mean difference 1.5 repeats, 95% CI 0.3, 2.7, p=0.013).

4.3.1.1 Differences between alternative test versions

There was a statistically significant difference between test forms for the HMGT (p=0.001) and so form was included as a covariate in any further analysis of this task. There was no evidence of differences between form for any other tests which had alternative versions.

Table 4-1 Demographic data for the cohort at baseline

	Control (N=20)	Premanifest (N=21)	Early HD (N=40)
Gender (M:F)	7:13	10:11	20:20
Age (year)[a]	44.9 (10.5)	37.2 (7.9)	48.5 (9.6)
Estimated premorbid IQ	106.2 (11.6)	103.2 (9.3)	105.3 (13.0)
Handedness (R:L)	19:1	20:1	36:4
CAG repeat length[b]	NA	42.2 (1.8), range 40 – 45	43.7 (2.4), range 40 – 50
Predicted years to onset[c]	NA	18.2 (7.1), range 9 – 35	NA
Disease duration (year)	NA	NA	4.1 (2.6)
UHDRS Motor[d]	1.1 (0.9)	3.6 (4.0)	28.9 (12.6)
UHDRS Independence	100 (0)	100 (0)	90.4 (9.6)
UHDRS TFC	13 (0)	13 (0)	10.9 (1.8)
BDI	5.6 (3.9)	6.8 (6.3)	9.3 (8.7)

NA: not applicable

Data are mean (SD) with the exception of gender and handedness; handedness was taken as the hand used to write with; UHDRS: motor is out of 124, higher score = more severely impaired; independence is a percentage, higher score = better function; Total Functional Capacity is out of 13, higher score = better function.

a PM<Control ($p=0.011$); PM<HD ($p<0.0001$)

b PM<HD ($p=0.013$)

c Onset was defined as a 60% chance of showing motor signs (Feigin *et al.* 2006) and predicted using the equation of Langbehn *et al.* (2004)

d HD<Control, HD<PM (both $p<0.0001$)

4.3.2 Clinical measures

Mean UHDRS motor score was higher in the early HD group than both controls and premanifest subjects (both $p<0.0001$, Table 4-1). Mean UHDRS motor score in the premanifest group was also higher than controls ($p=0.007$). The HD group was also clearly scoring below ceiling at the UHDRS independence scale and TFC.

Most subjects scored within the non-depressed range on the Beck Depression Inventory (BDI) (18 controls (90%), 15 premanifest (71%) and 24 early HD subjects (67%)), with 20 subjects classified as depressed in total (two controls (10%), six premanifest (29%) and 12 early HD subjects (33%)). Although the proportion of controls who were depressed was slightly smaller than that in the other two groups this difference was not statistically significant ($p=0.15$).

4.3.3 Brain volumes

After correcting for head size and adjusting for differences due to age the mean brain volume in the HD group was smaller (9%) than that in controls ($p<0.0001$, see Table 4-2) and the premanifest group ($p<0.0001$). The mean volume in the premanifest group was smaller (3%) than that in controls ($p=0.028$). The difference between clinical stage 1 and 2 symptomatic subjects was not statistically significant (data not shown).

Table 4-2 Mean (SD) baseline brain volumes, standardised to mean control TIV, with differences with and without adjustment for age

	Control (N=20)	Premanifest (N=21)	Early HD (N=40)	Difference (Premanifest - Control) Crude	Difference (Premanifest - Control) Adjusted	Difference (Early HD - Control) Crude	Difference (Early HD - Control) Adjusted
Brain volume (ml)	1185 (34)	1175 (47)	1062 (58)	-9.6 (p=0.45)	-27.2 (p=0.028)	-122.7 (p<0.0001)	-114.4 (p<0.0001)
95% CI	1169, 1201	1154, 1197	1044, 1081	-34.7, 15.5	-51.5, -3.0	-146.3, -99.2	-133.6, -95.1

Negative differences indicate smaller volume in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age; adjusted differences are the differences having adjusted for age by including it as a covariate in the linear regression models.

4.3.4 Cognitive performance

Group differences at baseline, adjusted for age and IQ, were converted to z-scores to facilitate comparisons between tests (by subtracting the control mean and dividing this by the control standard deviation). These are shown in Figure 4-2. Raw and age- and IQ-adjusted differences between groups are shown in Appendix 4.

4.3.4.1 UHDRS and other executive and psychomotor tasks

The HD group was statistically significantly worse than controls at category fluency ($p<0.0001$), phonemic fluency ($p<0.0001$), Stroop colour reading ($p<0.0001$), word reading ($p<0.0001$) and interference ($p<0.0001$), the SDMT ($p<0.0001$), the HMGT ($p=0.007$) and "A" cancellation ($p<0.0001$). The difference in UHDRS cognitive score (the sum of the three Stroop subtests, phonemic fluency and the SDMT) was also statistically significant ($p<0.0001$). The HD group was also slower than controls at TMT A and B and the difference between TMT B and A (all $p<0.05$), although not the ratio between these two scores. The HD group was also significantly worse than the premanifest group at all these measures.

There was no evidence of impairment in the premanifest group relative to controls although performance was approaching HD level at phonemic fluency (see Figure 4-2).

Very few subjects made errors on TMT A (one control (5%), five premanifest (24%), three early HD subjects (8%)), and all of these made just one error, with the exception of one premanifest subject who made two. The number of errors did not differ significantly between groups. On TMT B eight controls (40%),

seven premanifest (33%) and 21 early HD subjects (57%) made at least one error, and the early HD group made significantly more errors than the premanifest group ($z=2.25$, $p=0.025$).

4.3.4.2 Memory

The HD group was worse than controls at HVLT immediate recall ($p=0.001$), delayed recall ($p<0.0001$), percent recalled ($p=0.01$), and discrimination ($p<0.0001$). This group also scored, on average, fewer true positives ($p<0.05$) and more false positives ($p<0.05$) in the recognition task. After adjusting for immediate recall score, the early HD group was still impaired at delayed recall ($p=0.007$) and also at recognition discrimination ($p=0.011$). However, on average these subjects were no longer making fewer true positives than controls after adjusting for immediate recall score ($p>0.05$) although they were still making more false positives ($p<0.05$) (note that as 95% bootstrapped confidence intervals were used, precise p values were not obtained for these and similar analyses).

This group was also worse than controls at digit span forwards ($p=0.002$) and backwards ($p<0.0001$), RMF ($p<0.0001$) and RMW ($p<0.05$) and RMT difference (the difference between the RMF and RMW score, $p=0.025$, i.e. they were relatively more impaired on visual compared with verbal memory than controls were). The HD group was also significantly worse than the premanifest group on all these measures except the percent recalled on the HVLT.

There was no evidence of impairment in premanifest subjects relative to controls although they were approaching HD level in percent recalled in the HVLT.

4.3.4.3 Naming

The HD group was worse at naming than the premanifest group ($p=0.005$) but there was no evidence that performance differed from controls. The premanifest group performed slightly better than controls ($p=0.049$). The HD group was slower at naming than controls ($p<0.0001$) and the premanifest group ($p=0.002$) and this difference remained after controlling for dysarthria by including the dysarthria score as a covariate (HD vs. controls, $p<0.0001$; HD vs. premanifest $p=0.005$).

The difference in naming latency between premanifest subjects and controls was not statistically significant.

4.3.4.4 Visuoperceptual

There was no evidence of differences between groups on the VOSP. The HD group was worse than controls at the Benton Facial Recognition Test ($p<0.0001$), and also worse than the premanifest group ($p<0.0001$).

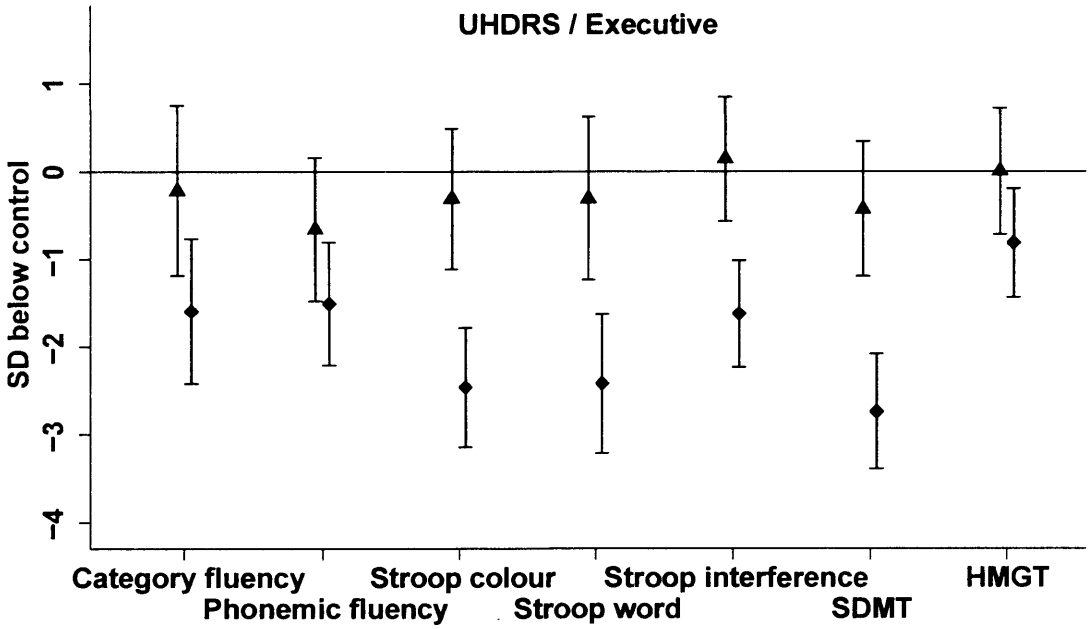
4.3.4.5 Reaction time

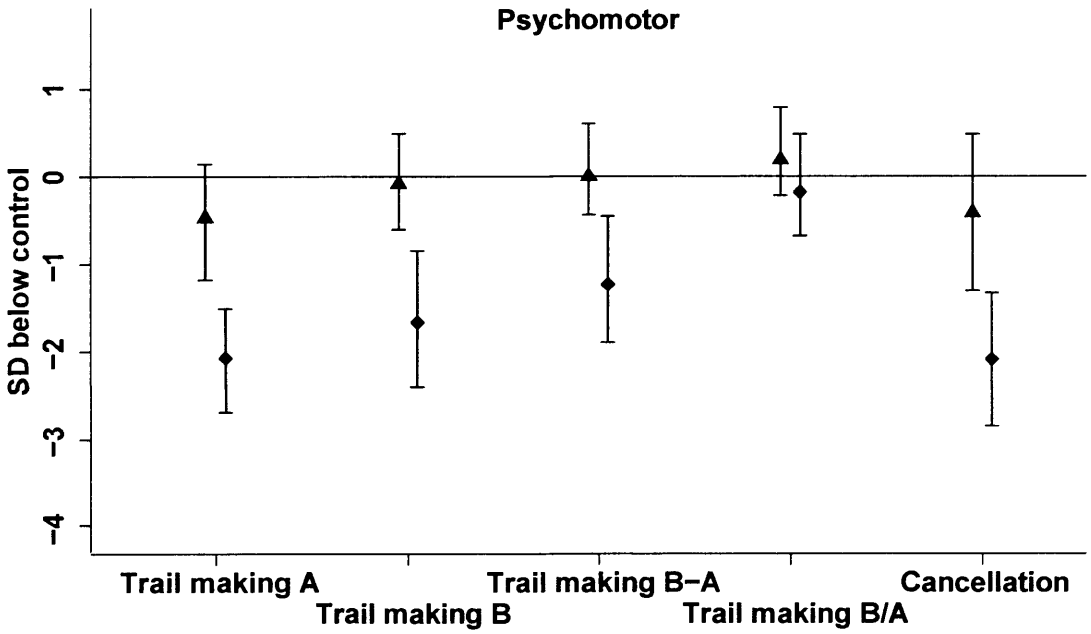
The HD group was slower than both controls and the premanifest group at all numbered button presses (both near and far, $p<0.05$) and there was no evidence of a difference between premanifest subjects and controls. The HD group was also slower than both controls and premanifest subjects at returning to the central blue button after near and far numbered button presses (both $p<0.05$) and the premanifest group was also slower than controls on these measures (both $p<0.05$).

Thirty-two subjects made no errors, 21 made one error (seven controls (35%), four premanifest (19%) and 10 early HD subjects (25%)) whilst 28 made two or more errors (eight controls (40%), seven premanifest (33%) and 13 early HD subjects (33%)). The number of errors did not differ statistically significantly between groups, and the maximum number of errors (and therefore discarded trials) per subject was six trials out of 80. Very few subjects made perseverative errors (two controls (10%), one premanifest (5%) and six early HD subjects (15%)) and the number of perseverative errors did not differ statistically significantly between groups.

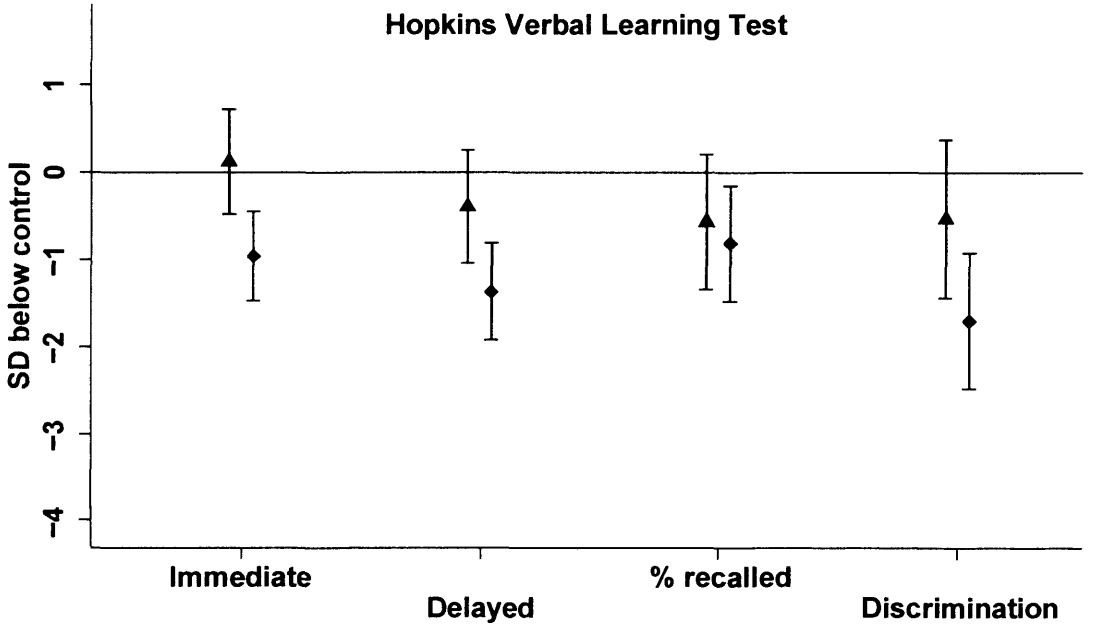
4.3.4.6 Emotion recognition

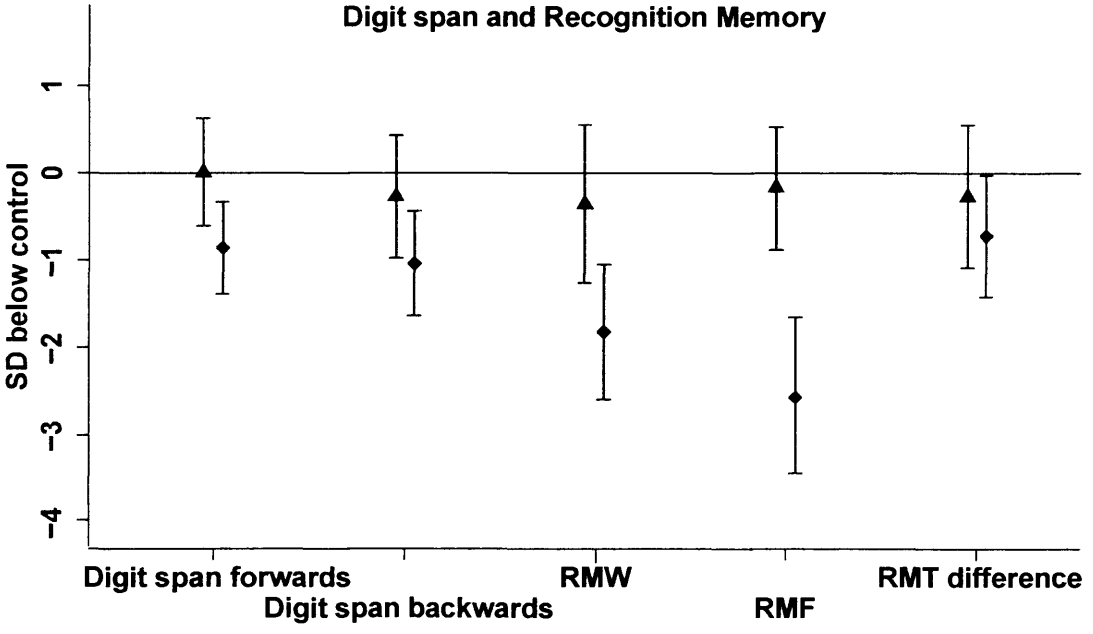
The HD group was significantly worse than controls and the premanifest group at total emotion recognition score ($p<0.05$). There was no evidence of a difference between premanifest subjects and controls. More detailed analysis of each emotion is reported in chapter 8.





▲ Premanifest ◆ Early HD





▲ Premanifest ◆ Early HD

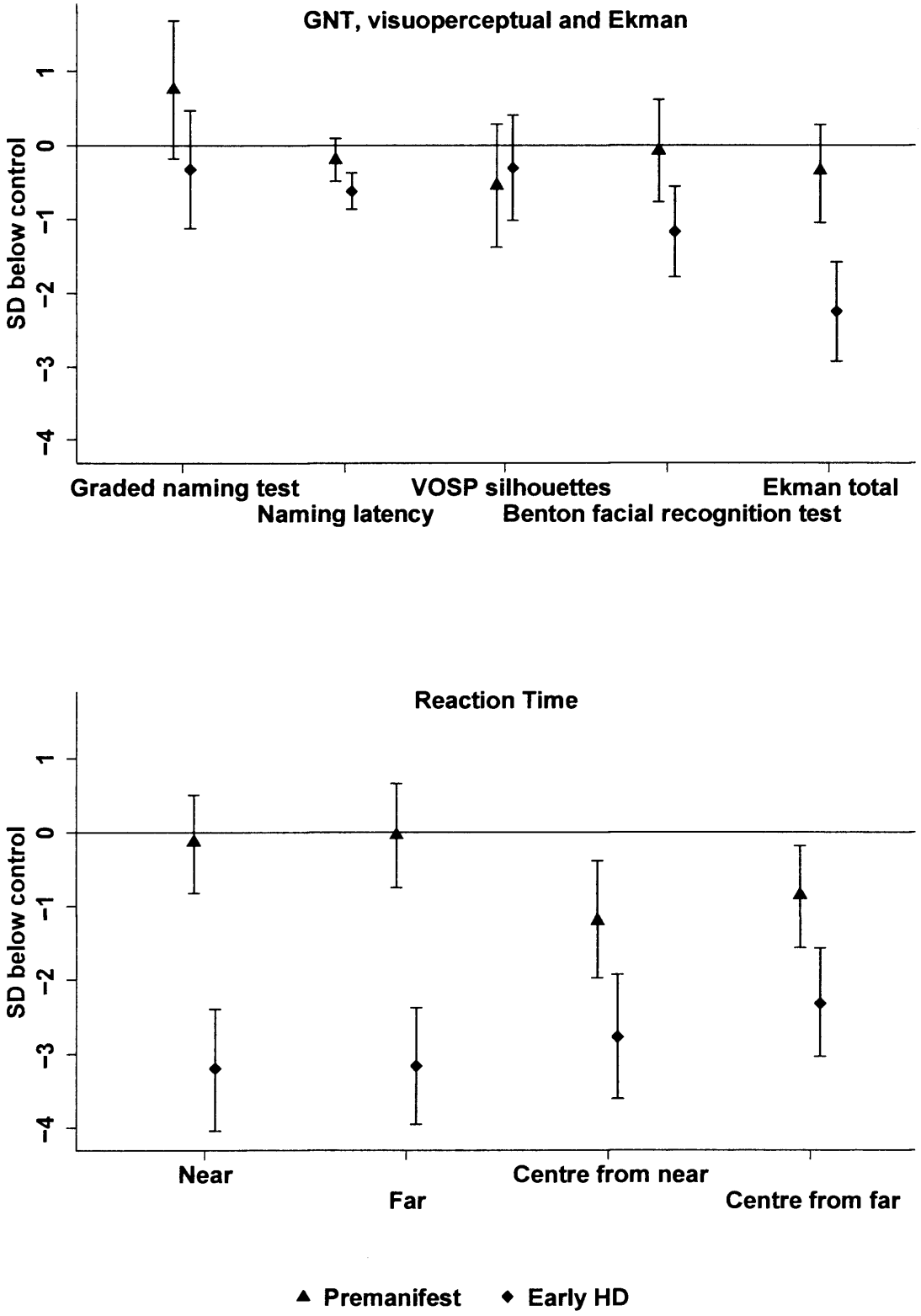


Figure 4-2 Age- and IQ-adjusted z-scores for the two patient groups relative to controls, with 95% confidence intervals

Overall the HD group showed evidence of impairment in all tasks other than naming and visuoperceptual skills. In contrast in the premanifest group levels of impairment were low and generally not statistically significant, although on some tasks it was noted that they were approaching HD levels, and on the RT task they showed evidence of some impairment.

4.3.5 Effect of motor impairment on psychomotor tasks in the early HD group

When motor score was controlled for by including it as a covariate the HD group was no longer impaired relative to controls at phonemic fluency ($p=0.27$), Stroop colour reading ($p=0.27$), word reading ($p=0.67$), trail-making A ($p>0.05$), cancellation ($p=0.68$) or the four RT measures (all $p>0.05$). However this group was still impaired relative to controls at Stroop interference ($p=0.041$), SDMT ($p=0.002$), and trail-making B ($p<0.05$), and the difference nearly reached statistical significance for category fluency ($p=0.063$).

4.3.6 Relationship between brain volume and clinical and cognitive measures in HD gene carriers

There was no evidence of a relationship between brain volume and UHDRS independence score, TFC or disease duration in the HD group alone.

There was a statistically significant effect of CAG repeat length on brain volume in the gene-carrying group as a whole (premanifest and early HD), after adjusting for age and group, such that an increase of CAG repeat length by one was associated with a 9.2ml decrease in brain volume (95% CI -17.5, -0.9, $p=0.03$, Figure 4-3). There was no evidence that the slope of this relationship differed between the premanifest and HD groups.

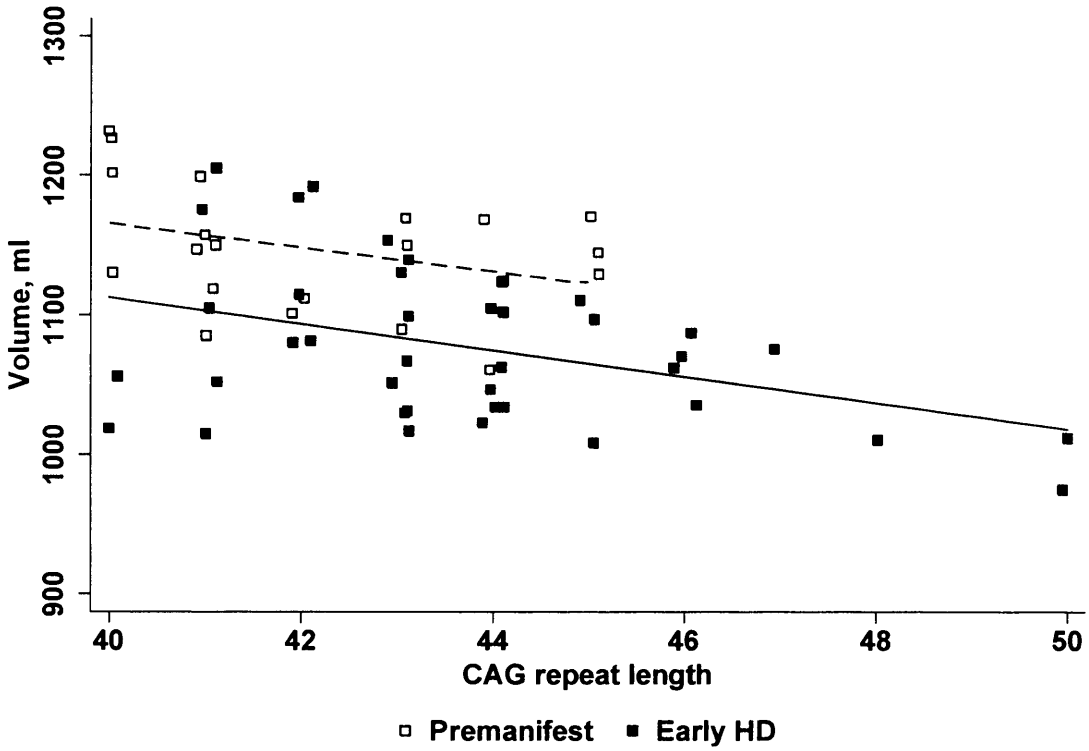


Figure 4-3 Relationship between brain volume (standardised to mean age in the whole cohort) and CAG repeat length in the premanifest and early HD groups

The relationship between volume and score differed between the premanifest and early HD groups for a number of tests. In the HD group, but not the premanifest group, smaller brain volume was associated with worse performance on UHDRS motor ($p=0.001$), HVLT discrimination ($p=0.001$), TMT A ($p<0.05$), Benton Facial Recognition ($p<0.0001$), Ekman emotion recognition ($p<0.05$), and near and far button press times (both $p<0.05$). In the premanifest group, but not the HD group, smaller volume was associated with worse performance on the GNT ($p=0.013$) (Table 4-3).

Table 4-3 Slope (95% confidence intervals) of the relationship between brain volume and score in the patient groups, when these slopes differed between groups, adjusted for age and estimated premorbid IQ

	Premanifest (N=21)	Early HD (N=40)
Test	Slope (points per ml)	Slope (points per ml)
UHDRS motor	-0.03 (-0.07, 0.01) p=0.08	-0.10 (-0.16, -0.04) p=0.001
HVLT discrimination	-0.01 (-0.02, 0.01) p=0.42	0.02 (0.01, 0.03) p=0.001
GNT	-0.03 (-0.06, -0.01) p=0.013	0.01 (-0.02, 0.03) p=0.76
TMT A (sec)	0.01 (-0.04, 0.07) p>0.05	-0.07 (-0.13, -0.02) p<0.05
Benton	-0.02 (-0.05, 0.02) p=0.35	0.04 (0.02, 0.07) p<0.0001
Ekman	-0.02 (-0.09, 0.07) p>0.05	0.10 (0.04, 0.17) p<0.05
Near button press (msec)	0.07 (-0.99, 1.31) p>0.05	-1.30 (-1.95, -0.64) p<0.05
Far button press (msec)	0.01 (-1.06, 1.41) p>0.05	-1.26 (-2.02, -0.51) p<0.05

Slope indicates the effect a 1ml increase in brain volume has on the respective test; units are always points per ml, unless the test is timed in which case units are specified in brackets; where 95% bootstrapped CIs were used precise p values were not obtained.

Overall there was a tendency for smaller brain volume to be associated with worse or slower scores at all tests. There was an effect of volume on score in all gene carriers (premanifest and HD) for HVLT immediate ($p=0.039$), RMF ($p=0.003$), RMW ($p<0.05$), phonemic fluency ($p=0.004$), SDMT ($p=0.001$), HMGT ($p=0.015$), TMT B ($p<0.05$), the difference between TMT A and B, and the ratio between them (both $p<0.05$) and VOSP silhouettes ($p=0.01$). In all cases smaller volume was associated with worse performance (Table 4-4).

Table 4-4 Slope (95% confidence intervals) of the relationship between brain volume and score in the 61 gene carriers, adjusted for age and estimated premorbid IQ

	HD gene carriers (Premanifest & HD)
Test	Slope (points per ml)
UHDRS cognitive	0.24 (-0.01, 0.49) p=0.05
Stroop colour	0.05 (-0.01, 0.11) p=0.09
Stroop word	0.02 (-0.08, 0.12) p=0.69
Stroop interference	0.01 (-0.03, 0.05) p=0.66
SDMT	0.08 (0.03, 0.12) p=0.001
Phonemic fluency	0.08 (0.03, 0.14) p=0.004
Category fluency	0.03 (-0.00, 0.06) p=0.059
HMGT	0.01 (0.003,0.02) p=0.015
TMT B (sec)	-0.41 (-0.67, -0.15) p<0.05
TMT B-A (sec)	-0.36 (-0.63, -0.12) p<0.05
TMT B/A (sec)	-0.007 (-0.01, -0.001) p<0.05
Cancelling As (sec)	-0.03 (-0.07, 0.01) p=0.15
HVLT immediate (/36)	0.03 (0.01, 0.05) p=0.039
HVLT delayed (/12)	0.01 (-0.01, 0.02) p=0.14
HVLT % recalled	0.01 (-0.09, 0.12) p=0.82
Digit span forwards (/21)	0.01 (-0.01, 0.02) p=0.57
Digit span backwards (/21)	0.01 (-0.01, 0.03) p=0.20
RMF (/50)	0.05 (0.02, 0.08) p=0.003
RMW (/50)	0.02 (0.01, 0.04) p<0.05
VOSP (/15)	0.02 (0.01, 0.03) p=0.01

Test	HD gene carriers (Premanifest & HD) Slope (points per ml)
Return from near button press (msec)	-1.48 (-2.27, -0.45) p<0.05
Return from far button press (msec)	-1.10 (-1.98, -0.12) p<0.05

Slope indicates the effect a 1ml increase in brain volume has on the respective test; units are always points per ml, unless the test is timed in which case units are specified in brackets.

4.3.7 Prediction of onset in the premanifest group

In general performance tended to decline slightly with estimated proximity to onset, although effects were very small and close to zero. There was a very slight tendency for higher motor score to be associated with proximity to onset (an increase of 2.1 points for each 0.1 increase in probability, 95% CI -2.7, 44.8, p=0.079). No other clinical, cognitive or MRI measures were associated with probability of motor onset.

4.3.8 Comparison of NART and the Spot the Word Test

The two measures were highly correlated (r=0.78, p<0.0001). The mean difference between IQ measured by the NART and the Spot the Word Test did not differ between groups, and hence mean (SD) IQ measures for the whole cohort are shown in Table 4-5, with the Spot the Word scores converted to have the same distribution as the NART. The mean difference between the two measures was statistically significant (p=0.0053) with the Spot the Word Test giving, on average, significantly lower estimates of IQ than the NART.

Table 4-5 Mean (SD) NART and Spot the Word scores for the whole cohort at baseline, with difference (95% confidence intervals)

	Mean (SD) (N=81)
NART	105.0 (11.7)
Spot the Word	102.2 (13.8)
Difference	2.7 (8.6)
95% CI	0.8, 4.6

The standard deviation of the differences was 8.6 and the limits of agreement were -14, 20. Variance was higher in the Spot the Word test (p=0.022) although this was driven by two outliers in the premanifest group and not representative of the data overall.

4.4 DISCUSSION

Deficits have been shown across cognitive domains in this cohort of early HD subjects, including immediate and delayed recall, recognition memory, face recognition and facial emotion recognition, as well as executive and psychomotor tasks. In addition, early HD subjects had increased naming latency although they were able to name similar numbers of items to controls. There was little evidence of impairment in the premanifest subjects on the same battery, other than increased reaction times at button pressing.

4.4.1 Cognitive deficits in the early HD cohort

The early HD cohort is fairly typical of many reported elsewhere, consisting of predominantly stage 1 and 2 subjects, with a mean disease duration of just over four years. Compared to the control sample their greatest deficits were mostly on

motor or psychomotor tasks such as the reaction time task, SDMT, Stroop colour and word reading, and Trail Making A, but interestingly they also showed similar levels of deficit on the RMW, and the Ekman emotion recognition test (based on the z scores of the group differences).

Early HD subjects showed a greater absolute difference between TMT A and B than controls, but the ratio between the two times did not differ between groups, suggesting that although the HD subjects were slower, they were not disproportionately slower for the more difficult task. HD subjects also showed relatively greater impairments at Stroop colour and word reading than Stroop interference, adding weight to suggestions that one of the primary deficits in HD is a loss of "automaticity" which manifests as an inability to perform relatively automated tasks (such as word reading) whilst tasks which naturally require more cognitive control (such as the Stroop interference task) are relatively less affected (Snowden *et al.* 2001).

In order to assess the contribution of motor ability to the impairments seen on these psychomotor tasks, group differences were adjusted for motor ability by including UHDRS motor score as a covariate. The HD group was no longer significantly impaired relative to controls at phonemic fluency, Stroop colour reading, word reading, trail-making A, cancellation or the four RT measures. However this group was still significantly impaired at Stroop interference, SDMT, and trail-making B, and the difference approached significance for category fluency. It seems likely that in the former tasks motor skills contribute significantly to performance but that in the latter poor motor skills alone are not sufficient to explain impaired performance, which may reflect problems with

higher-level cognitive functions. It is notable that trail-making A performance (joining numbers in sequence) was unimpaired after motor score was taken into account, whereas trail-making B performance, with its more cognitively-demanding switching component, was still impaired. Both phonemic and category fluency depend on phonemic switching ability which has been shown to decrease with increasing motor score in HD (Rich *et al.* 1999; Ho *et al.* 2002). However, adjusting for motor score had a greater effect on the phonemic fluency than the category fluency impairment, which suggests that of the two phonemic fluency may rely more heavily on motor ability, perhaps because of the difficulties associated with rapid production of words beginning with the same phonemes.

Patients with HD were also impaired at the relatively novel test of homophone meaning generation (Warrington 2000). This test requires multiple switches between concepts and is untimed. It has been shown to be sensitive to frontal lobe dysfunction and has an advantage in being less subject to ceiling effects than some tests in this domain. This was demonstrated by the wide range of scores in all three subject groups, although the premanifest subjects tended to perform as well as, and in some cases better than, controls. Follow up of this population will demonstrate whether subtle drop off in performance can be detected as the premanifest subjects approach onset.

Although memory skills were generally impaired, the largest effects were seen for recognition memory (both words and faces and HVLT discrimination). Free recall of words (both immediate and delayed) and digit span forwards and backwards were also impaired, but to a lesser extent. This appears to contrast

with some earlier reports that the main impairment in HD is one of retrieval (as opposed to storage) and that this is reflected in better recognition than free recall scores (see Craufurd and Snowden 2002) and agrees with a recent meta-analysis showing that in general the effect sizes for group differences in recall were no different to those for recognition (Montoya *et al.* 2006). However, typically in recognition tests controls are performing at or near ceiling, and their variance is therefore reduced, which can artificially inflate the effect of group differences. An examination of the raw data suggested that this might be the case for the RMW and HVLT discrimination score, as well as HVLT delayed recall, at which a number of subjects, but mainly controls, scored at ceiling.

Relative to controls, early HD subjects were impaired at each part of the HVLT; they learnt fewer words in the early trials, recalled proportionately fewer after a delay, and recall was not improved to control levels by testing recognition. The recognition score reported here is the number of true positives minus false positives in an auditory "yes / no" recognition test, meaning that subjects could correctly recognise all 12 targets but subsequently be penalised for false recognition of distractor items. Although everyone tended to score quite highly on this task, the early HD group tended to get fewer true positives and more false positives than controls, suggesting that their impairment arose from both a failure to recognise target items as well as misclassification of distractors.

However, when recall and recognition performance were adjusted for the number of words recalled originally (an index of encoding) there was no longer a statistically significant difference between the number of true positives recognised in the early HD and control groups, although this difference remained

for delayed recall and for the number of false positives recognised. This confirms theories that, *given their level of encoding*, early HD subjects can correctly recognise as many target stimuli as controls. This also bears out Lang *et al.*'s (2000a) account, which suggests that if false positives are taken into account in recognition test scores then there is little to distinguish recall and recognition performance in HD. Early HD subjects were still making more false positives than controls, and this contributed to the difference in recognition discrimination score even when immediate recall score was taken into account.

The RMT is different, in that each target is presented with a distractor and the subject must choose one, meaning that the total recognition score takes errors into account. Again, recognition of words and faces was impaired, with a relatively greater impairment shown for faces, although again this might be partly due to the reduced variance in controls, a number of whom scored at ceiling on this task. It is hard to compare the RMT with the HVLT since the paradigms are different, but again these results demonstrate an impairment in a recognition task that includes false positive scores, and that these impairments cover auditory verbal stimuli, and visual verbal and facial stimuli.

It has been suggested that in HD a failure to organise information or to use strategies during encoding and recall, rather than an inability to store information *per se*, underlies the impairments seen on memory tasks, and these results seem to confirm this. The data presented here also suggest that subjects with early HD also show increased levels of false positive recognition, which may also be related to disruption to prefrontal regions thought to be involved in monitoring retrieval in recognition tasks (Henson *et al.* 2000). Further work with this cohort

could investigate whether this pattern changes with the introduction of strategies during learning; even if this were the case, a paradigm such as the HVLT, in which true and false target stimuli are presented separately, might still be sensitive to impairment in early HD if subjects have difficulties judging their level of confidence in target stimuli.

The early HD group was not impaired at a very basic test of visuoperceptual skills, but was impaired at facial recognition and facial emotion recognition, a pattern of findings that is common in early disease (Jacobs *et al.* 1995; de Boo *et al.* 1997; Lawrence *et al.* 2000; Milders *et al.* 2003). The facial emotion recognition deficit is discussed in more detail in chapter 8.

There was no evidence that naming was impaired in the early HD cohort, which is typical of the early stages, although naming deficits have been detected in subjects as the disease progresses (Hodges *et al.* 1991; Craufurd and Snowden 2002). Errors were too few to be analysed by category but HD subject errors were predominantly either "don't know" errors or perceptual, e.g. mistaking the handcuffs for binoculars or the tassel for a broom and then correcting this after prompting. However there were also some ambiguous semantic / perceptual errors (mistaking the buoy for a sandcastle) and semantic / phonemic errors (labelling the handcuffs "cufflinks"). Despite not showing significant impairment at naming, early HD subjects were significantly slower than controls. The majority (twenty-seven) of the 40 HD subjects scored zero on the dysarthria subscale of the UHDRS ("normal") with 11 subjects rated as "unclear but no need to repeat" and the remaining two as "must repeat". Even when dysarthria scores were adjusted for in the latency analysis group differences remained,

suggesting that slower naming was at least partly due to cognitive slowing rather than an articulatory deficit. Reaction times for production of phonemes with different articulatory requirements are unimpaired in early HD (Ludlow *et al.* 1987) which again reinforces the theory that the slower naming times seen here were not due to motor speech problems. Cognitive slowing has been demonstrated previously in HD using an executive function task (Hanes 1996) but has not been shown before in confrontation naming. It would be interesting to follow the premanifest cohort to investigate whether similar slowing could be demonstrated prior to motor onset.

4.4.2 Cognitive deficits in the premanifest cohort

The 21 premanifest subjects were on average 18 years from estimated motor onset, and only one of them was less than 10 years from onset so as a group these subjects were unlikely to show differences from control performance. However, although in this study there was very little evidence of impairment on any of the tests used, others have reported that even when performance on these tests is still within clinically-normal limits, deficits can be detected with sufficiently large subject numbers, and that performance tends to decline with proximity to estimated motor onset (Snowden *et al.* 2002; Paulsen *et al.* 2006a; Solomon *et al.* 2007; Stout *et al.* 2007; Verny *et al.* 2007). It also seems likely that decline in all modalities in premanifest subjects is very slight until within the last 5-10 years prior to estimated onset, at which point the rate of decline increases and appears relatively linear (Paulsen *et al.* 2007). As most of the cohort in this study were estimated to be more than 10 years from onset, they were likely to be undergoing, at most, only very subtle decline. Although there was a slight trend

for greater UHDRS motor scores to be associated with proximity to onset, effects on other tasks were very small and non-significant.

Despite this, in most cases the premanifest mean was worse than controls suggesting that levels of performance might already be beginning to decline. The group was significantly impaired at returning to press a central blue button in the RT task which supports other findings that motor abnormalities may be detectable even in those more than 10 years from predicted onset (Stout *et al.* 2007) although in addition to a button-pressing task these authors also found TMT B to be impaired, which was not found in the current study.

The RT task had two components, pressing one of four numbered buttons in response to an unpredictable on-screen stimulus, and then returning to press the central blue button in response to a predictable on-screen stimulus. The early HD group was impaired at both these aspects, which is consistent with their slowed motor abilities. The premanifest group was only impaired at the predictable part of the task, so they were no slower than controls in their reaction to the number, but failed to take advantage of the knowledge that the central button was always pressed after a number trial. The ratio of time to press unpredictable to predictable buttons in controls was 1:0.4, whilst in premanifest and early HD subjects it was 1:0.5 and 1:0.6 respectively.

A relatively greater decline in simple motor tasks compared with more cognitively demanding ones has been demonstrated in manifest HD, and it was suggested that an impairment in implementing and executing automatic tasks, relative to those under effortful control, might underlie this (Snowden *et al.* 2001). A similar deficit in acquiring a repetitive motor response might explain

the pattern of results seen here, and this fits well with similar deficits in motor sequence learning found elsewhere (Feigin *et al.* 2006). In particular it has been shown that even under explicit learning conditions premanifest subjects are impaired at spatial accuracy and modulating movement time, relative to controls learning the same task, whereas there was no difference between groups when targets appeared in a random order (Ghilardi *et al.* 2008). However the cross-sectional results presented here also suggest that early HD subjects, although slowed, are no slower than premanifest at the predictable part of the task. This finding therefore needs to be replicated in a larger group, nearer to estimated onset and followed longitudinally. If confirmed, tasks of this nature might be suitable for detecting very early psychomotor deficits in this group.

It is also important to appreciate that the estimate of years to onset is likely to be somewhat conservative, given that here, as elsewhere in the HD literature, onset was defined as a 60% chance of showing motor signs, and that the existing model does not take into account other variables which have been suggested to modify age at onset (e.g. parental age at onset, paternal vs. maternal transmission) (Langbehn *et al.* 2004). Whilst this measure allows comparison of premanifest cohorts between studies, and it is likely that the majority of the current cohort are relatively far from onset, a more accurate representation of their current status will only be possible retrospectively, after longitudinal follow-up confirms their true time of motor onset.

4.4.3 Brain volumes and associations with cognitive performance

Whole-brain volumes were similar to those reported in chapter 3, with the HD group on average 114ml smaller than the control group. Differences in

methodology make it hard to compare absolute volumes between studies but this difference equates to a 9% loss in HD relative to controls, which is similar to the differences found in other early HD populations, ~9% by Kassubek *et al.* (2004a), and between 10% (grey matter) and 17% (white matter) by Fennema-Notestine *et al.* (2004). Others have found reductions of smaller magnitudes which were not significant or only reached significance in more advanced subjects (Aylward *et al.* 1998; Beglinger *et al.* 2005), but this may be partly to do with the smaller sample sizes (and corresponding lack of statistical power) in these studies. As the striatum is less than 20ml in healthy adults (Aylward *et al.* 2004) some (the majority) of the loss in the HD group must be attributable to extra-striatal atrophy.

Total brain volume in the premanifest group was also decreased relative to the controls. Although smaller brain volumes have been reported elsewhere in premanifest subjects these differences were not statistically significant (Reading *et al.* 2005; Paulsen *et al.* 2006b). Brain volume in all gene carriers was associated with CAG repeat length (having adjusted for age). Neither age nor CAG repeat length was associated with disease duration in the early HD cohort so this was not simply an effect of those with longer CAG repeat lengths having had the disease for longer at the time of entry to the study. It may be that those with higher CAG repeat lengths have a faster rate of progression but this needs to be confirmed longitudinally.

In the gene carriers as a whole smaller brain volume was indicative of worse performance at memory tests, as well as some executive function tasks and a visuospatial task, suggesting that performance at these tasks falls off gradually

with increasing brain loss. In contrast smaller brain volume was associated with worse performance at the more overt motor tasks, as well as HVLT discrimination, and face and emotion recognition, in the HD group alone, suggesting that performance on these tasks is not strongly related to atrophy until after a certain point has been reached (in this case motor onset). Smaller brain volume was associated with worse performance on the GNT in the premanifest, but not the early HD group, which is hard to interpret given that this was one of the few tasks on which neither group showed impairment relative to controls.

This suggests that whole-brain volume is a sensitive measure, despite the fact that it encompasses regions of the brain that are likely to be unaffected as well as those that are. Structure-function relationships cannot be inferred from associations with whole-brain volume but it seems likely that atrophy in extrastriatal, as well as striatal regions, must underlie these associations, given the magnitude of the differences in brain volume seen here. Whole-brain volume has previously been found to be a good predictor of cognitive function (Harris *et al.* 1996) and as mentioned in chapter 1, there is increasing evidence that cortical and white matter atrophy is widespread quite early on in the disease (Rosas *et al.* 2005; Paulsen *et al.* 2006b; Rosas *et al.* 2008).

4.4.4 Other comments

The Spot the Word test correlated well with the NART as an estimate of premorbid intelligence although underestimated IQ slightly across all groups (assuming that the NART, as the more established of the two, is giving accurate measures of IQ). However as this bias was the same for both HD subjects and controls, Spot the Word score could still be used as a covariate to adjust for

differences due to premorbid IQ. The limits of agreement between the two measures were quite wide, however, meaning that the Spot the Word Test could give an estimate that was approximately one standard deviation over or under the NART estimate. Given that the NART is well-established as an estimate of premorbid IQ, it may be preferable to use the NART (at least in early HD) if IQ needs to be estimated accurately, rather than for use as a nuisance covariate.

It had been suggested that the Spot the Word might be less susceptible to problems related to HD such as dysarthria, which might unfairly penalise subjects on the NART test which requires words to be pronounced out loud. The ANART has been found to decrease very slightly with proximity to onset in premanifest subjects (Solomon *et al.* 2007) and so this work is followed up in chapter 9 which looks at the stability of both tests over time.

Although this study used a range of cognitive tasks in a relatively large HD cohort it was still subject to a number of limitations. The battery, whilst extensive, could not cover all domains in detail and thus these findings may not fully describe the overall nature of the deficits seen in HD. Few tests of visuoperceptual function and naming were used, in part because these skills tend to be relatively less affected in the early disease stages and I wanted to focus on skills that were likely to show promise as markers of change using standard neuropsychometric tests. It is likely that deficits in these domains could have been detected in this cohort with more sensitive tests, as others have shown previously (Hodges *et al.* 1991; Lawrence *et al.* 2000).

Whilst the early HD cohort was relatively large, and in terms of CAG repeat length and disease duration, relatively homogeneous, the premanifest subjects

covered a wide range of estimated years to onset, and only one was thought to be within 10 years of clinical diagnosis (although as discussed it must not be forgotten that this measure is only an estimate). Subjects were recruited according to the criteria in section 2.1.1, i.e., they needed to have a copy of the abnormal gene without having been diagnosed as showing unequivocal motor signs. There was no attempt to recruit subjects at a particular proximity to onset, and CAG repeat sizes were not known until after recruitment was complete. The small size of this cohort, coupled with the distance from onset, meant it was unlikely that effects would be large enough to detect and in most domains this was the case. This reflects the nature of the cohort, and does not contradict the wide-ranging deficits which have been demonstrated in substantially larger, closer-to-onset cohorts (Paulsen *et al.* 2007).

4.5 CONCLUSION

This study demonstrated a wide range of impairments in early HD covering psychomotor, executive function, memory and emotion recognition, but sparing visuoperceptual and naming skills. Of note, deficits in recognition memory and emotion recognition were as great as those shown on some executive and psychomotor tasks, and increased naming latency was shown, in the absence of a naming deficit. In some, but not all tasks with a motor component, performance was explained by the subjects' poor motor ability.

The premanifest group was on the whole very far from onset, at a stage where little impairment was expected to be evident. However impairment could be detected on some tasks, in particular the RT task for which results suggested that

these subjects had relatively more difficulty acquiring a predictable motor skill than controls.

Whole-brain volume, despite being a measure that encompasses regions which are atrophying as well as regions which are not, was associated with CAG repeat length and performance on a number of cognitive tasks, suggesting that it may prove to be a sensitive marker of disease progression.

Work in the following chapters builds on these results, using both VBM (a whole-brain approach) and caudate volumetry to elucidate further the patterns of atrophy and structure-function relationships in this cohort. This cross-sectional study also served as a baseline assessment, and longitudinal follow-up of these subjects is reported in later chapters.

5 CAUDATE VOLUME AND ITS RELATIONSHIP WITH COGNITION IN HUNTINGTON'S DISEASE

5.1 INTRODUCTION

Caudate atrophy is one of the earliest and most striking features of Huntington's disease and has been demonstrated in premanifest subjects at around 15 years prior to estimated motor onset (Aylward *et al.* 2004; Paulsen *et al.* 2006b), but not in subjects who were on average more than 20 years from onset (Paulsen *et al.* 2006b). However, it is important to note that different estimates of onset have been used, including one that takes parental age at onset into account (Aylward *et al.* 2004) and the probabilistic survival model of Langbehn *et al.* (2004), and in some cases it is not clear what probability of onset was used to determine years to onset (Paulsen *et al.* 2006b). Caudate atrophy has also been shown to progress with advancing disease (Aylward *et al.* 1997; Aylward *et al.* 2004).

Recent work suggests that fronto-striatal circuits are important for integrating sensorimotor, associative and affective information in order to learn and modulate behaviours in response to the environment (Costa 2007; Samejima and Doya 2007). The striatum is thought to play a role in goal-directed behaviour and skill-learning by associating actions with rewards and hence reinforcing particular stimulus – action – outcome associations (Poldrack 2002; Delgado 2007). With increasing learning these associations become habitual or "automatic" and this is associated with a shift in activation from the cortex to the striatum (Saling and Phillips 2007) and from medial to more dorsal striatal regions (Balleine *et al.* 2007; Costa 2007).

Parallel and largely functionally-segregated cortico-striatal circuits have been described (Alexander *et al.* 1986), which are thought to underlie motor, oculomotor, associative (executive function) and emotional processes in humans (Penney, Jr. and Young 1983; Cummings 1993; Bonelli and Cummings 2007). More recently it has been suggested that the circuits are not completely separated and this explains why motor, cognitive and emotional symptoms are seen in HD, even though in the early stages damage is primarily to the putamen and dorsomedial caudate (motor and associative circuits) rather than the ventral striatum (thought to be part of the emotional circuit) (Joel 2001).

Many studies in HD assumed that caudate or putamen atrophy caused the early cognitive signs of the disease, although few included both neuropsychometry and imaging in order to investigate this directly. There is some evidence that reduced caudate volume is associated with psychomotor, executive and visuospatial dysfunction in HD (Bamford *et al.* 1989; Starkstein *et al.* 1992; Peinemann *et al.* 2005; Douaud *et al.* 2006) although not all these studies tested function in other domains and those that did reduced cognitive skills to sets of factors making it hard to tease out specific structure-function correlates.

Manual segmentation is considered the "gold standard" method for quantifying volumes on MRI. However, although generally considered to be more accurate than automated methods, manual methods can be open to bias and measurement error. Techniques have been developed to minimise these effects, such as aligning scans in a standard orientation prior to segmentation, and blinding the segmentor to laterality. Reproducibility can also be aided by using interactive thresholding and arbitrary cut-offs. This study demonstrates a protocol which

aimed to minimise potential operator bias and maximise inter- and intra-rater reproducibility.

The work in this chapter therefore aimed to measure caudate volume reliably, to examine whether caudate atrophy could be detected in the current cohort of premanifest HD subjects, and to investigate the clinical and cognitive correlates of this atrophy.

5.2 METHODS

5.2.1 Subjects

Subjects were the same 81 as those reported in chapter 4, see section 4.2.1 for details.

5.2.2 Assessments

Subjects underwent the same assessments described in chapter 4, sections 4.2.2.1, 4.2.2.2, and 4.2.2.3. That is, they underwent clinical and cognitive assessments and had 1.5T MRI scans.

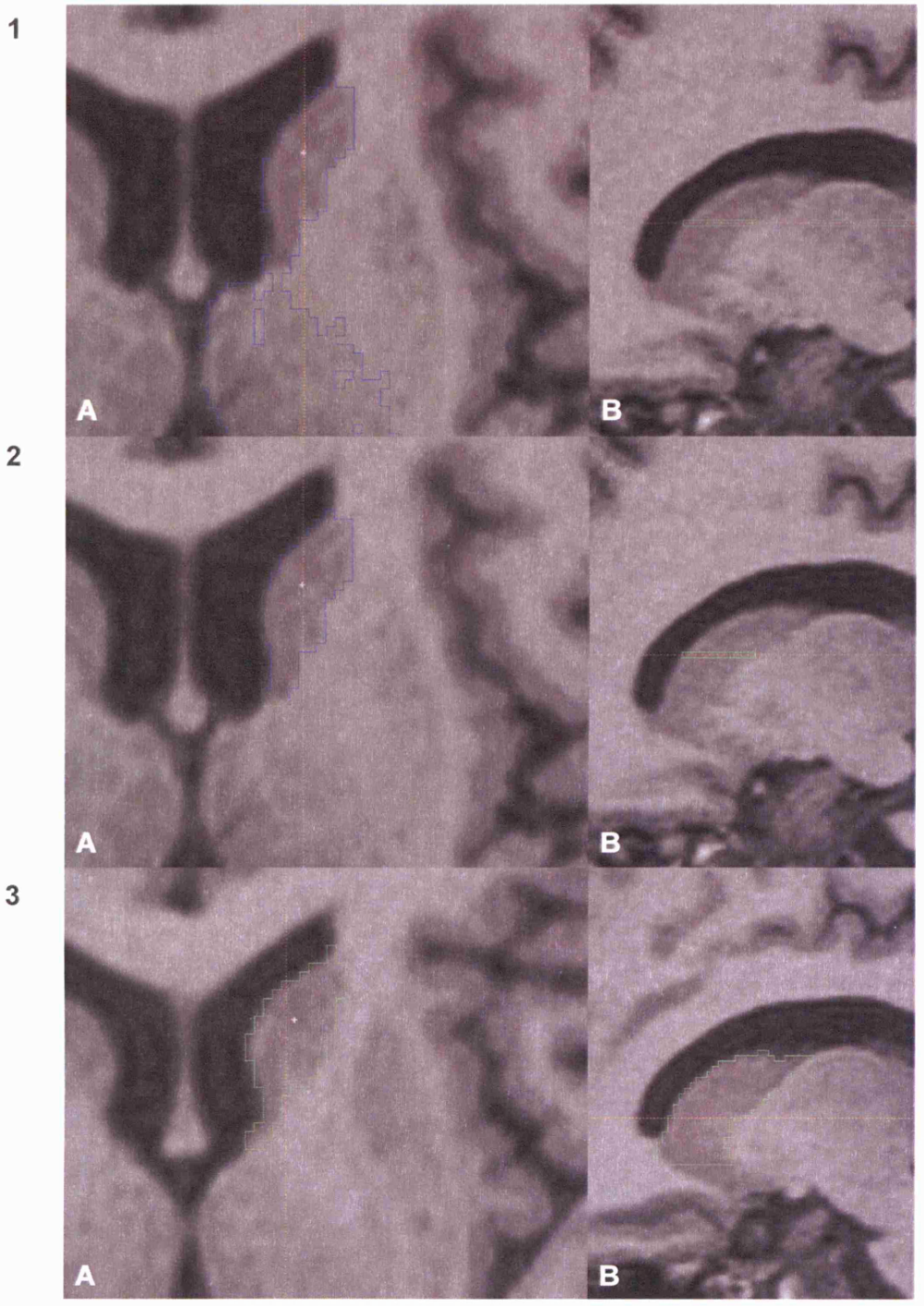
5.2.3 Caudate segmentation

Scans were segmented using MIDAS (see section 2.5.3.1 and Figure 5-1). Whole brains were segmented as in chapter 4. Whole brains were then registered into MNI (standard) space using a six degrees of freedom (dof) algorithm (using translations and rotations). This resulted in all whole brains being oriented in the same standard space to facilitate consistent segmentation.

Each image was then flipped across the mid-sagittal plane, in order to produce two mirror-image scans. The caudate was always segmented on the right-hand

side of the image. The head and body of the caudate were included with the medial border defined by the lateral ventricle and the lateral border by the internal capsule. Anteriorly the caudate was included as soon as it was clearly visible next to the ventricle coronally, and posteriorly and superiorly the body was included for as long as was visible. Elsewhere arbitrary cut-offs were imposed to improve reproducibility: inferiorly segmentation stopped below the last slice in which the ventricle was visible, and the nucleus accumbens was excluded on coronal view, in line with the inferior edge of the ventricle if no border was clear.

Caudates were segmented axially from the most superior slice on which the structure was visible next to the lateral border of the ventricle, until the slice below the most inferior view of the ventricle. Intensity threshold constraints were used for initial delineation of the caudate boundary, using lower and upper thresholds of 69% and 110% of mean brain intensity to exclude CSF and WM respectively. Two orthogonal views were available to the operator who manually edited the boundary in those regions for which the thresholds were not sufficient (Figure 5-1). The initial region was then edited in coronal view to ensure that the nucleus accumbens was excluded. All caudates were segmented by two raters whose intra- and inter-rater reproducibility (intraclass correlation coefficients) was >0.99 (mean absolute difference ~4% of mean caudate volume).



Caudates were segmented on the right of the image in axial view, with sagittal (or coronal) view available for reference and later editing. Panel 1 shows initial seed placement and boundaries defined by thresholds; panel 2 shows the same slice after manual editing of the region that spilled outside the caudate; panel 3 shows the region after further segmenting; A shows axial view, B shows sagittal view; (continued on following page).

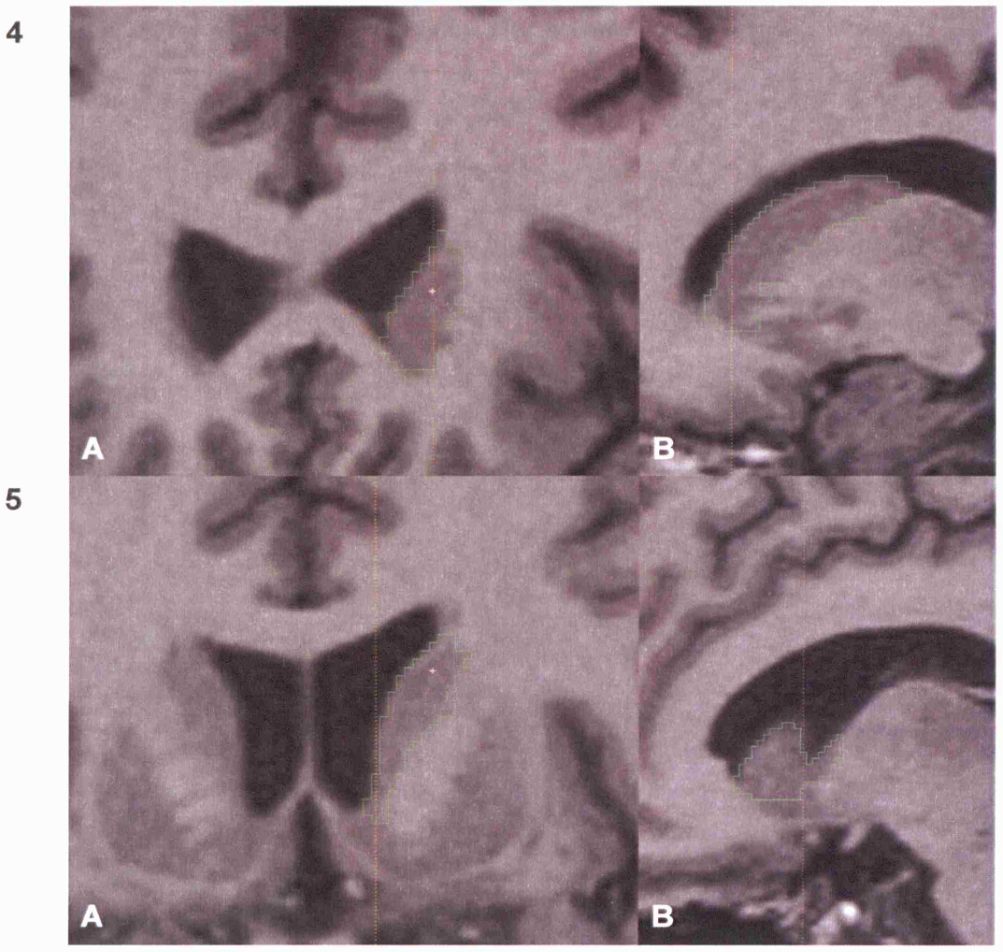


Figure 5-1 Screenshots of MIDAS software being used to segment the caudate nucleus

Panel 4 shows the region in coronal view (A) and sagittal view (B); panel 5 shows where the nucleus accumbens has been excluded on coronal view.

5.2.4 Statistical analysis

To correct for differences in head size baseline caudate volumes were standardised for TIV as outlined in section 3.2.3.2. Where caudate volume is mentioned in the results it always refers to volume standardised to mean control TIV.

5.2.4.1 Effect of adjusting for head size on gender differences

As it has previously been shown that adjusting caudate volumes for the mid-skull area (an estimate of head size) reduced but did not eliminate differences due to gender (Raz *et al.* 1995), gender differences in caudate volume in controls were assessed both before and after standardising to mean control TIV, by regressing each measure of caudate volume on gender, with robust standard errors, and adjusting for age.

In the following analyses the effects of variables which can independently affect brain volume (age) and cognitive performance (age and estimated premorbid IQ) were controlled for by including them as covariates where necessary.

5.2.4.2 Group differences in caudate volume

In order to investigate whether differences between left and right caudate volumes differed between groups, a linear regression model was used to compare the ratio of left to right caudate volumes between groups, adjusting for age.

A linear regression model was then used to compare caudate volumes (both left and right caudates) between groups, with robust standard errors to allow variances to differ between groups, and adjusting for age. In order to investigate whether the association between volume and variables of interest differed between the left and the right caudate, variables of interest were regressed on a difference measure (left caudate - right caudate).

5.2.4.3 Associations with clinical and cognitive variables

Separate linear regression models were used to look at the associations between caudate volume and estimated probability of onset (Langbehn *et al.* 2004) in the

premanifest group, and UHDRS TFC, independence score and disease duration in the early HD group, adjusting for age.

As in chapter 4, separate linear regression models (with robust standard errors) relating either CAG repeat length, cognitive scores or UHDRS motor score, to total caudate volume, group and their interaction, were used to investigate whether the association between these variables and caudate volume differed between the premanifest and HD groups. In situations where there were significant interactions the slope of the association in each group was examined to see if, in either group, the association was significantly different from zero. Secondly, for variables for which there was no evidence of an interaction, linear regression models were used to look for an association between volume and score in all gene carriers, having adjusted for differences due to group. Age and estimated premorbid IQ were included as covariates in all cognitive analyses.

The models described above were also used for data which were not Normally distributed, but where Normality assumptions were violated 95% bias-corrected bootstrap confidence intervals with 1000 replicates were used; these data were the RMW, trail-making and reaction times, and the Ekman emotion recognition score.

5.3 RESULTS

As subjects were the same as those reported in chapter 4, demographic data can be found in Table 4-1, page 135.

5.3.1 *Effect of adjusting for head size on gender differences*

Prior to adjustment for TIV males had, on average, larger total caudate volume than females (mean difference 1.1ml, 95% CI 0.6, 1.7ml, p<0.0001, after adjusting for age, see Table 5-1). After adjustment for TIV this difference was reduced and no longer statistically significant (mean difference 0.2ml, 95% CI -0.3, 0.8ml, p=0.34, after adjusting for age, see Table 5-1). This mean difference was therefore reduced from ~14% to <3% of caudate volume (~8ml).

Table 5-1 Mean (SD) caudate volumes in male and female controls, before and after adjusting for TIV, with differences (95% confidence intervals)

	Male	Female	Difference (95% CI)
Caudate volume (ml)[a]	(N=7)	(N=13)	
Raw	8.6 (0.5)	7.5 (0.6)	1.1 (0.6, 1.7)
TIV-adjusted	8.1 (0.5)	7.8 (0.6)	0.2 (-0.3, 0.8)

a Volumes presented here are age-adjusted

Caudate volumes presented in the remainder of this chapter have all been standardised to mean control TIV.

5.3.2 *Group differences in caudate volume*

There was no evidence of a difference in the ratio of left to right caudate volumes between groups, after adjusting for age (all p>0.09). A *post hoc* paired t-test on all 81 subjects showed that the right caudate was on average significantly larger than the left (mean difference 0.08 ml, 95% CI 0.04, 0.11 ml, p<0.0001).

The premanifest group had smaller caudates than controls (left: p=0.005; right: p=0.001. ~12% smaller for both left and right) and the early HD group had smaller caudates than both controls and the premanifest group (vs. controls, left

and right: both $p<0.0001$, ~40% smaller for both left and right; vs. premanifest, left and right: both $p<0.0001$, ~30% smaller for both left and right), adjusted for age (see Table 5-2, Figure 5-2).

Table 5-2 Mean (SD) left and right caudate volumes in each group, standardised to mean control TIV, and differences with and without adjustment for age

	Control (N=20)	Premanifest (N=21)	Early HD (N=40)	Difference (Premanifest - Control)		Difference (Early HD - Control)	
				Crude	Adjusted	Crude	Adjusted
Left (ml)	3.88 (0.30)	3.47 (0.60)	2.27 (0.50)	-0.41 p=0.006	-0.43 p=0.005	-1.61 p<0.0001	-1.60 p<0.0001
95% CI	3.74, 4.02	3.20, 3.74	2.11, 2.43	-0.70, -0.12	-0.73, -0.14	-1.81, -1.40	-1.80, -1.39
Right (ml)	4.04 (0.26)	3.54 (0.61)	2.31 (0.47)	-0.50 p=0.001	-0.51 p=0.001	-1.73 p<0.0001	-1.72 p<0.0001
95% CI	3.92, 4.16	3.26, 3.82	2.16, 2.46	-0.78, -0.21	-0.80, -0.22	-1.92, -1.54	-1.92, -1.53

Negative differences indicate smaller volume in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age; adjusted differences are the differences having adjusted for age by including it as a covariate in the linear regression models.

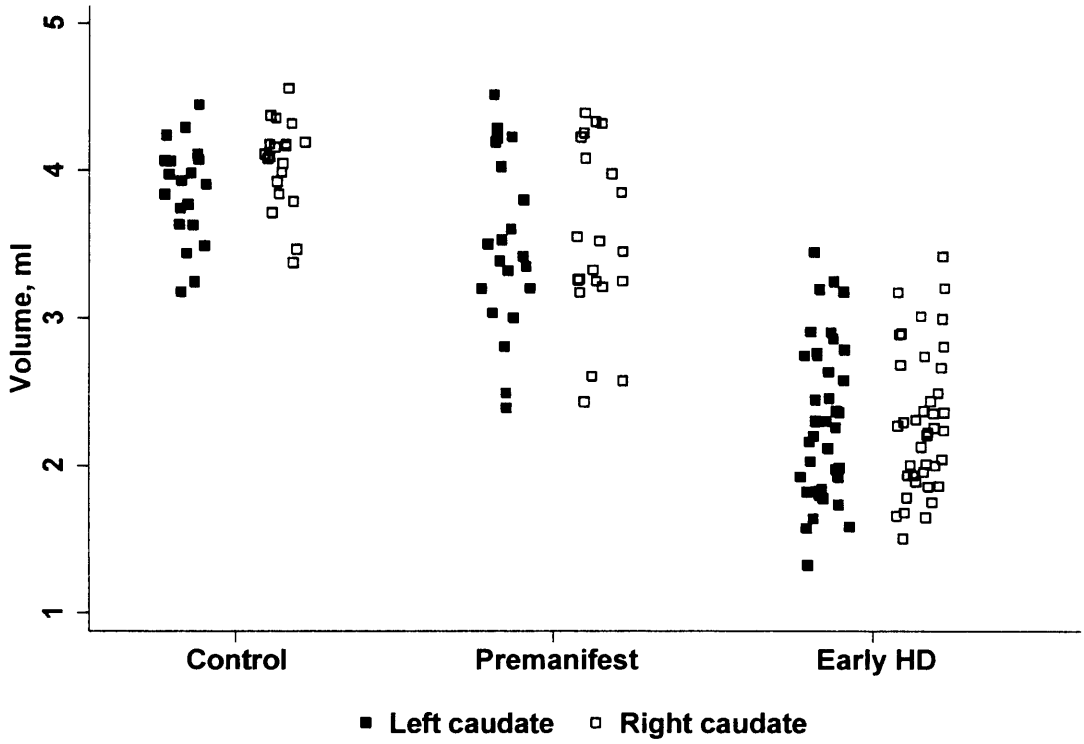


Figure 5-2 Left and right caudate volumes in each group, standardised to mean age in the whole cohort

There was no evidence that the relationship between volume and other variables differed statistically significantly between the left and right caudate. This was true for CAG repeat length and cognitive scores (all gene carriers), probability of onset (premanifest group), disease duration, UHDRS motor score, total functional capacity and independence score (early HD group). Hence for all further analyses the left and right caudate volumes were combined.

5.3.3 Associations with clinical variables

In the 61 gene carriers CAG repeat length was associated with total caudate volume, after adjusting for age and group, such that an increase in CAG repeat length of 1 repeat was associated with a decrease in total caudate volume of 0.3ml (95% CI 0.1, 0.5ml, $p<0.0001$) (Figure 5-3). There was no evidence that the slope of this association differed between the premanifest and early HD

groups. The addition of disease duration to this model (looking at early HD subjects only) had little effect on the relationship (an increase of 1 CAG was associated with a 0.3ml decrease in total caudate volume, 95% CI 0.1, 0.4ml, p=0.013).

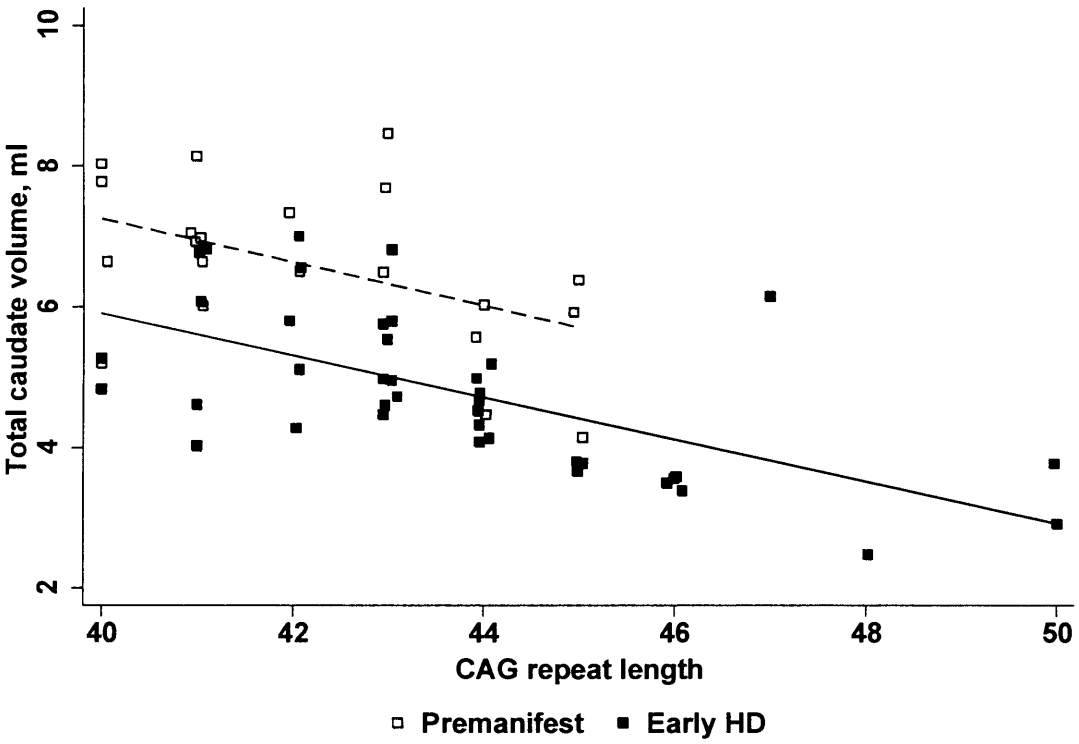


Figure 5-3 Relationship between caudate volume (standardised to mean age in the whole cohort) and CAG repeat length in the premanifest and early HD groups

In the 21 premanifest subjects there was a tendency for greater probability of onset to be associated with smaller total caudate volume, such that an increase in probability of onset by 0.1 was associated with a 0.6ml decrease in total caudate volume (95% CI -0.4, 13.2ml, p=0.065), after adjusting for age. Nine out of 21 premanifest subjects had caudate volumes smaller than the smallest control, and six of them were more than 11 years from estimated motor onset (range 12–23 years).

There was no evidence of a relationship between total caudate volume and disease duration, independence score, or total functional capacity in the early HD group (all $p>0.49$).

The slope of the relationship between UHDRS motor score and caudate volume differed between groups ($p=0.017$). In the premanifest group a 1ml decrease in volume was associated with a 2.0 point increase in motor score (95% CI 0.3, 3.6 points, $p=0.019$) and in the early HD group a 1ml decrease in volume was associated with a 7.4 point increase in score (95% CI 3.9, 10.9 points, $p<0.0001$), after adjusting for age.

5.3.4 Associations with cognitive variables

The relationship between cognitive score and caudate volume was significantly different between the premanifest and early HD groups for TMT A. There was no evidence of an association between score and volume in the premanifest group ($p>0.05$), but in the early HD group a 1ml decrease in caudate volume was associated with a 4.4 second increase in TMT A time (95% CI 1.5, 7.8 sec, $p<0.05$).

For all other variables there was no evidence that the slope of the relationship between score and caudate volume differed between groups. Regression coefficients demonstrating the effect of a 1ml increase in caudate volume on scores in all gene carriers (N=61) are presented in Table 5-3. In general, smaller caudate volume was associated with lower scores or slower times at tests. Smaller caudate volume was significantly associated with worse UHDRS cognitive score ($p=0.003$), Stroop colour ($p=0.003$), Stroop word ($p=0.006$), SDMT ($p=0.005$), phonemic fluency ($p=0.011$), RMF ($p=0.007$), VOSP

($p=0.002$), Benton facial recognition ($p=0.033$), Ekman facial recognition of surprise, disgust and anger (all $p<0.05$) and time to press numbered buttons (all $p<0.05$). There was also a tendency for smaller volume to be significantly associated with worse scores on category fluency ($p=0.09$), HMGT ($p=0.07$) and cancellation ($p=0.06$). There was no evidence of an association between score and caudate volume for Stroop interference, TMT B, all the subtests of the HVLT, digit span forwards and backwards, RMW, the GNT, Ekman facial recognition of happiness, sadness and fear, and returning to press the central blue button (all $p>0.05$).

Table 5-3 Slope (95% confidence intervals) of the relationship between caudate volume and score in the 61 gene carriers, adjusted for age and estimated premorbid IQ

	HD gene carriers (Premanifest & HD)
Test	Slope
UHDRS cognitive	17.0 (6.2, 27.9) p=0.003
Stroop colour	4.7 (1.7, 7.7) p=0.003
Stroop word	6.1 (1.8, 10.3) p=0.006
Stroop interference	0.3 (-2.2, 2.7) p=0.83
SDMT	3.2 (1.0, 5.5) p=0.005
Phonemic fluency	2.8 (0.7, 4.8) p=0.011
Category fluency	1.1 (-0.2, 2.4) p=0.09
HMGT	0.5 (-0.03, 1.0) p=0.07
TMT B (sec)	-4.6 (-13.0, 4.7) p>0.05
TMT B-A (sec)	-2.4 (-10.5, 7.1) p>0.05
TMT B/A (sec)	0.03 (-0.2, 0.3) p>0.05
Cancelling As (sec)	-1.6 (-3.4, 0.1) p=0.06
HVLT immediate	0.8 (-0.5, 2.1) p=0.22
HVLT delayed	0.2 (-0.3, 0.7) p=0.48
HVLT % recalled	0.2 (-4.9, 5.4) p=0.92
HVLT discrimination	0.3 (-0.1, 0.7) p=0.13
Digit span forwards	0.1 (-0.6, 0.9) p=0.71
Digit span backwards	0.9 (-0.2, 2.1) p=0.10
RMF	2.2 (0.6, 3.8) p=0.007
RMW	0.2 (-0.6, 1.0) p>0.05

Test	HD gene carriers (Premanifest & HD) Slope
GNT	-0.1 (-1.1, 0.8) p=0.78
VOSP	0.8 (0.3, 1.3) p=0.002
Benton	1.1 (0.1, 2.1) p=0.033
Ekman happiness	-0.01 (-0.08, 0.04) p>0.05
Ekman sadness	0.02 (-0.1, 0.2) p>0.05
Ekman surprise	0.2 (0.004, 0.3), p<0.05
Ekman disgust	0.4 (0.2, 0.6) p<0.05
Ekman anger	0.4 (0.1, 0.6) p<0.05
Ekman fear	0.1 (-0.2, 0.4) p>0.05
Far button press (msec)	-55.2 (-88.0, -30.5) p<0.05
Near button press (msec)	-45.8 (-74.0, -18.2) p<0.05
Return from far button press (msec)	-35.4 (-76.6, 9.3) p>0.05
Return from near button press (msec)	-28.1 (-71.4, 13.3) p>0.05

Slope indicates the effect a 1ml increase in caudate volume has on the respective test; units are always points per ml, unless the test is timed in which case units are specified in brackets; where 95% bootstrapped CIs were used precise p values were not obtained.

5.4 DISCUSSION

5.4.1 Caudate volume in HD

Caudate volume was significantly reduced in both early HD and premanifest subjects in this cohort. As with brain volumes it is difficult to compare absolute measures between studies, because of the different scanning parameters and segmentation protocols employed, and whether or not volumes are presented adjusted for head size. However in general where the head and body of the

caudate have been measured they are reported to be about 60% of control volume in early HD (Rosas *et al.* 2003; Beglinger *et al.* 2005) which is very similar to the difference seen here, although one study in early HD found volumes to be less than 40% of control (Rosas *et al.* 2001). This seems more similar to volumes seen in advanced patients (Fennema-Notestine *et al.* 2004; Mascalchi *et al.* 2004), which is approaching the levels seen at post-mortem (de la Monte *et al.* 1988; Mann *et al.* 1993).

Caudate volume in this cohort of premanifest subjects, an average of 18 years from onset, was about 88% of that in controls, compared with 78% in a nearer-to-onset cohort (on average 11 years from onset) (Paulsen *et al.* 2006b), 85% (subjects on average 14 years from onset) (Aylward *et al.* 2004) and 93% (subjects on average 23 years from onset) (Paulsen *et al.* 2006b), the latter difference not reaching significance. Together these results suggests a decrease in volume with proximity to onset and this was true of the current cohort although there was some variability in this, with some nearer-to-onset subjects with volumes in the control range, and some farther-from-onset subjects with volumes in the early HD range. Overall these results suggest that it may be possible to detect reduced caudate volume slightly earlier than previously thought, and certainly well before most cognitive or standard clinical measures can detect abnormalities (see chapter 4).

In this study a rightward asymmetry was found in both controls and HD subjects, with no evidence that this differed between groups. This contrasts with the findings of Rosas *et al.* (2001) who found a leftward bias in controls, and that this was reversed in subjects with HD. However, of the few studies in HD that

have reported left and right caudate volumes, others have consistently found right caudates to be larger both in controls and premanifest and HD subjects (Mascalchi *et al.* 2004; Reading *et al.* 2005) although in the former case the difference was not significant and in the latter it was not tested. One other study found no difference between right and left caudate volumes (in controls and early HD) although absolute volumes were not reported (Ruocco *et al.* 2006). Findings in healthy subjects have been varied (Raz *et al.* 1995) but the majority of studies have demonstrated a rightward bias in healthy caudate volumes (Peterson *et al.* 1993; Raz *et al.* 1995). Therefore, although this work replicates that of Rosas *et al.* (2001) in finding a rightward asymmetry in caudate volumes in premanifest and early HD, this appears to parallel the relationship found in the healthy caudate and therefore does not support the suggestion that there is a reversal of asymmetry in HD. One study has reported an increase in rightward asymmetry (left-sided loss) in the caudate in HD relative to controls, (Mühlau *et al.* 2007a) but this has not been replicated. This issue merits further investigation as it may change as the disease progresses (something suggested by Mühlau *et al.*). In this study all caudates were segmented on the left, with segmentors blinded to laterality so neither the protocol nor the segmentor could have biased the laterality findings.

Caudate volume was associated with CAG repeat length, which has been found previously (Rosas *et al.* 2001; Ruocco *et al.* 2006) but not consistently (Aylward *et al.* 1996; Harris *et al.* 1999) although the latter two studies had relatively small numbers and may have lacked power. As with whole-brains, this relationship could not be explained by longer disease duration or age at onset which suggests that caudate atrophy rate might vary with CAG repeat length, and there is some

evidence that this might be the case (Aylward *et al.* 1997; Aylward *et al.* 2000). It is also important to consider the inherent uncertainty surrounding age at onset (and therefore disease duration), and the fact that onset is defined as unequivocal motor signs (as opposed to signs in other domains) all of which may make disease duration a somewhat unreliable measure.

Caudate volume was also associated with UHDRS motor score in both patient groups, although a decrease in caudate volume of 1ml had a greater effect on motor score in the early HD than the premanifest group, perhaps reflecting the fact that a 1ml difference in the early HD represents on average a greater proportional loss of volume. However this may also be due to the non-linearity of the UHDRS motor score, with premanifest subjects likely to score at ceiling on this subtest. Although others have found a similar relationship between motor score and caudate volume, the association tends to be stronger with putamen volume (Harris *et al.* 1992; Harris *et al.* 1996; Harris *et al.* 1999). Evidence suggests that the putamen, rather than the caudate, is involved in fronto-striatal motor circuits (Alexander *et al.* 1986; Joel 2001) so the finding that motor score is associated with caudate volume may in part reflect the fact that the size of the two structures is highly correlated. Caudate volume was not associated with the TFC or Independence score. An association with TFC has been reported previously (Aylward *et al.* 2003; Rosas *et al.* 2003) in similar subjects, so this relationship needs further investigation. Inter-rater differences between studies may contribute to the disparate findings.

5.4.2 Relationship between caudate volume and cognition

In general smaller caudate volume was associated with worse performance on psychomotor, fluency or visuoperceptual tasks, rather than verbal memory tasks, naming, and tasks with a cognitive switching component (Stroop interference, TMT B, category fluency and HMGT). This was true across all gene carriers as there tended to be little difference in the slope of the association between the premanifest and early HD groups. The exception to this was TMT A for which there was a relationship with caudate volume in the early HD, but not the premanifest group, which may be because performance on this task remains relatively unaffected until atrophy of the caudate (and other structures) is more extensive.

Few volumetric studies in HD have included measures of cognitive performance. One study employing a large battery of tests identified a set of factors and found that caudate volume was most strongly associated with complex psychomotor tasks (including Stroop, Digit symbol, TMT A and B), followed by visuoperceptual tasks, and was less strongly associated with verbal learning, (Bamford *et al.* 1989). This fits well with the findings reported here, and although some studies have failed to replicate specific associations this may be due to lack of power as they had smaller sample sizes (Harris *et al.* 1996; Beglinger *et al.* 2005). Studies taking a whole-brain approach have also found that striatal volume was associated with psychomotor or executive function, although associations with other domains were not tested (Kassubek *et al.* 2005; Peinemann *et al.* 2005; Douaud *et al.* 2006).

Many of the tasks for which an association was seen between caudate volume and performance are complex; most have an obvious motor and visual scanning component, both of which might depend to some extent on striatal integrity (Alexander *et al.* 1986), as well as requiring either a learnt skill (such as word-reading or number sequencing) or including a learning component (e.g., the SDMT, in which acquiring symbol-digit associations would improve task performance).

Degeneration of the dorsolateral striatum (putamen and dorsolateral caudate) is likely to result in deficits in well-learned (habitual or "automatic") tasks (Balleine *et al.* 2007; Costa 2007). Deficits in implementing an automatic response program are thought to underlie at least some of the impairments seen in HD (Snowden *et al.* 2001) and the current results support this. Performance on Stroop colour, word and TMT A, all of which rely on relatively well-learnt skills, was worse in subjects with smaller caudate volumes, whilst performance on TMT B and Stroop interference, although impaired in this cohort, was not associated with caudate volume. The Stroop and trail-making tasks are largely equated for motor and scanning components, so if the deficits seen on the tasks were solely due to motor and scanning difficulties, one would expect the association between caudate volume and performance to be the same for all parts of the tasks. This was not the case, and the main difference was that subjects with smaller volumes performed worse on the aspects of the tasks which required learnt routines. The Stroop interference condition is thought to rely on prefrontal regions to inhibit attentional processes to the task-irrelevant dimension (Banich *et al.* 2000; Harrison *et al.* 2005) and TMT B similarly may rely on dorsolateral prefrontal cortex (Zakzanis *et al.* 2005). The results presented here suggest that

whilst the caudate is related to performing habits or learning novel associations, aspects of executive tasks such as switching between two skills (number sequencing and letter sequencing) or over-riding a learnt skill (Stroop interference) are not related to the caudate as strongly. In fact, it may be that dorsal striatal damage and the associated deficit in performing habitual skills are advantageous in the Stroop interference task but that oculomotor and motor difficulties also contribute to slow performance. It would be interesting to investigate whether there are qualitative differences in the errors made depending on whether lesions affect the dorsal striatum or other regions; those with deficits at carrying out over-learnt skills might be less likely to make perseverative errors that those with, for example, lesions to the more medial fronto-striatal circuits who tend to be unable to inhibit habitual responses. Degeneration of the striatum is also likely to affect novel learning of stimulus-response associations (Poldrack 2002; Balleine *et al.* 2007), which could affect performance on tasks such as the SDMT.

Phonemic and category fluency are thought to depend on the ability to produce clusters of related words, and to switch between clusters (Troyer *et al.* 1998); similarly the HMGT is likely to require the ability to switch between meanings for a given word (Warrington 2000). Switching and clustering have been shown to rely on prefrontal regions and temporal lobe regions respectively (Troyer *et al.* 1998; Hirshorn and Thompson-Schill 2006). There is increasing evidence that switching ability is relevant for semantic as well as phonemic fluency (Hirshorn and Thompson-Schill 2006) and a deficit in switching between phonemic subcategories has been associated with the deterioration seen over time on both tasks in HD (Ho *et al.* 2002). In the current study there was a relatively strong

association between caudate volume and phonemic fluency (a decline of 2.8 words per ml decrease in volume) whereas volume was less strongly (and not significantly) related to category fluency and HMGT. It is likely that frontal dysfunction contributes to the impairments seen across all three tasks in this cohort, but it may be that phonemic fluency relies more heavily on phonemic switching (which correlates with motor ability (Ho *et al.* 2002)) than the other two tasks, and hence shows a stronger association with striatal volume.

It is also important to consider that performance on all tasks with an overt motor component (such as psychomotor tasks, fluency, cancellation and reaction time used here) may also be associated with putamen volume, which is likely to be correlated with caudate volume. It is probable that, as for the UHDRS motor score discussed above, loss of putamen volume (and associated motor deficits) underlies at least some aspects of impaired performance on these tasks.

Performance on tasks which had a clear visuoperceptual component (the RMF, VOSP, Benton facial recognition test and several emotions from the Ekman emotion recognition battery) was also associated with caudate volume. Unlike those discussed above, these tasks do not depend on over-learnt (habitual) skills, are untimed, and do not require the acquisition of novel associations. Poor performance here may be related to disruption to the oculomotor circuit, which includes the central body of the caudate nucleus (Alexander *et al.* 1986; Cummings 1993). Three of the four tasks used facial stimuli, and visual scanning is known to impact on facial learning and recognition (Henderson *et al.* 2005) and facial emotion recognition (Wong *et al.* 2005). There is evidence that eye movements are impaired in premanifest and early HD (Golding *et al.* 2006;

Blekher *et al.* 2006), and although visuoperceptual deficits are not considered one of the primary problems early in the disease, subtle impairments have been noted (Jacobs *et al.* 1995; Lawrence *et al.* 2000). The suggestion that abnormal eye movements contribute to poor visuoperceptual processing also means that it may be possible to improve performance, at least in the early stages of the disease, by directing subjects' attention to specific regions of the stimulus (Adolphs *et al.* 2005). This is discussed further in chapter 8, which investigates facial emotion recognition in more detail.

In contrast, performance on verbal measures of memory, both orally and visually presented, was not associated with caudate volume, although all these skills were impaired in the early HD group (see chapter 4). As discussed in the previous chapter, memory deficits in HD are thought to be due to poor encoding and retrieval, rather than problems with storage *per se*. This is difficult to explain in terms of striatal dysfunction, and may be due to damage, or reduced input, to associated prefrontal cortical structures (Henson *et al.* 2000; Wolf *et al.* 2007; Wolf *et al.* 2008). Whilst psychomotor task performance declines gradually in premanifest subjects (during which time striatal damage also seems to be progressive) there is evidence that memory performance deteriorates more suddenly around the time of clinical onset (Snowden *et al.* 2002). Recent work suggests that there is functional change in the dorsolateral prefrontal cortex (associated with working memory performance) in premanifest subjects, in the absence of volumetric differences (Wolf *et al.* 2007). It may be that frontal dysfunction (and atrophy) in HD is secondary to reduced input from striato-thalamic circuits and hence that it, and the associated behavioural deficits, become apparent some time after striatal atrophy can be detected.

There was no evidence that the relationship between caudate volume and score differed between the left and the right caudate. As the volumes of the two are highly correlated this would be difficult to demonstrate, and it may be that there is some functional lateralisation which could be investigated further using techniques such as fMRI.

5.4.3 Technical considerations

As mentioned above, it is hard to compare absolute caudate volumes in the literature because of the range of scan parameters and segmentation protocols. However there are also more subtle differences which have a bearing on the variability and reproducibility of the measure.

Some studies do not align images into standard space, or blind operators to laterality, and most use manual outlining without the aid of intensity thresholds. In the current study attempts were made to minimise differences between images (for example due to alignment or laterality) by registering images into standard space, so that segmentation was done on similar views of each image. In order to reduce operator bias, images were flipped so that both the left and the right caudate were segmented on the right side of the scan, with segmentor blinded to laterality. In order to improve reproducibility segmentation used interactive thresholding and arbitrary cut-offs, with minimal manual editing. Intraclass correlation coefficients for the two segmentors in the current study compare favourably with those reported elsewhere (0.99 compared with around 0.97 reported by Aylward *et al.* (2000; 2003)). The use of arbitrary cut-offs means that the method does not provide an estimate of total caudate volume, but nevertheless this measure does include the head and body of the caudate, whilst it

does not encompass all of the tail. The result of using thresholding and arbitrary cut-offs is a standardised protocol that is reliable and reproducible, which is important for large-scale trials with a number of operators or measurements done over several time periods.

Volumes were also standardised for head size, which adjusts both for differences in head size between groups (Whitwell *et al.* 2001) and differences in gender. Studies which neither match exactly for gender nor adjust for head size are likely to over- or underestimate group differences in caudate volume depending on the relative proportions of gender in each group. It is also of note that although adjusting for head size (as was done here) eliminated difference due to gender, adjusting for intra-skull area did not (Raz *et al.* 1995). Whilst ideally these methods need to be compared on the same dataset, intra-skull area is likely to be a more noisy estimate of head size than a volume measure, because it will be unduly influenced by head shape in the single plane in which it is measured, and hence it may be important to estimate head size more accurately in order to fully eliminate differences.

5.5 CONCLUSIONS

A protocol using standard-space-aligned, left-right flipped, and TIV-corrected volumes appears to give a measure of loss that compares favourably with methods in the literature, and suggests it is a relatively reliable and unbiased means of assessing caudate atrophy.

This work has shown that reduced caudate volume can be detected in premanifest gene carriers an average of 18 years prior to estimated motor onset. Individuals with higher CAG repeat lengths tended to have smaller volumes regardless of

disease duration or age, suggesting that higher CAG repeat length may be associated with higher rates of caudate atrophy.

Associations between caudate volume and cognition support the role of the striatum in learnt (automatic) skill performance, goal-directed behaviour and in motor, oculomotor and psychomotor circuits, and suggest that striatal damage may not underlie deficits seen in memory tasks, or tasks under more cognitive control, such as those involving a switching component. These findings also suggest that although impairments seen in HD are often described as "executive function" (problems with planning, set-shifting, directing attention), there may be a qualitative difference between impairments on tasks such as the SDMT and phonemic fluency, which were related to caudate volume, and TMT B and Stroop interference, which were not. As with any structure–function correlations, future work should focus on establishing the causality of these relationships with functional imaging, and also on investigating the relative contributions of frontal and striatal sub-regions in more detail.

6 VOXEL-BASED MORPHOMETRY: THE EFFECTS OF SOFTWARE VERSION, TEMPLATES, AND CONFOUNDING VARIABLES

6.1 INTRODUCTION

As discussed in chapter 1 (section 1.2.2), the technique of VBM is increasingly used in the study of neurodegeneration, and is an alternative, complementary approach to ROI methods because it is relatively automated and can be applied across the whole-brain and thus does not require *a priori* hypotheses about particular regions of interest. However VBM findings in HD are mixed. Whilst in some cases this might be due to the heterogeneity of the disease, it is possible that this is also caused by methodological or technical differences between studies, rather than reflecting real differences between populations. Despite being automated to some extent, a large number of parameters in the VBM process can be set by the operator and some processing steps are optional. In addition, statistical models vary widely.

The initial aims of this chapter were to maximise the accuracy of the SPM2 normalisation and segmentation algorithms on the dataset used in this thesis, and to investigate the effects of some potential confounding variables. This led on to further work examining different estimates of intracranial volume (to adjust for head size differences). Finally, this chapter reviews previous VBM studies in HD in order to compare techniques and identify methodological differences which might partly explain discrepant findings in the literature.

6.1.1 Maximising the accuracy of normalisation and segmentation in SPM

The images on which the statistical analysis is performed in VBM are preprocessed to align them in the same space and to segment the different tissue

types. The assumption of the analysis is that across a group of images all voxels at a given spatial location represent the same tissue type, and the same part of the brain (Ashburner and Friston 2000). Inferences about the location of atrophy and the anatomical substructures affected depend to some extent on the accuracy of these preprocessing steps. There is also a circularity in the preprocessing steps in that segmentation would be better following accurate registration but registration would be improved by being performed after segmentation. Although other preprocessing steps can alter the interpretation of findings, as can the way in which differences are then modelled, misclassification of tissue or poor inter-subject alignment could introduce basic errors into what is reported.

The standard VBM procedure normalises whole brains to a template whole-brain image prior to segmentation using prior probability maps, both of which are smoothed averages of a large number of images from healthy volunteers (e.g. the MNI-152 template, which is an average of 152 brains provided by the Montreal Neurological Institute). However it was noted that both alignment and segmentation could be improved by normalising each tissue type to a tissue-specific template image and by basing segmentation on study-specific prior probability maps (Good *et al.* 2001b). The authors called this an "optimised" method and since then many others have used aspects of it, although not all stick to precisely what was outlined in the original paper, meaning that the term "optimised VBM" actually covers a range of non-standard normalisation and segmentation options (compare e.g., Karas *et al.* 2003; Douaud *et al.* 2006; Wolf *et al.* 2007).

SPM software is updated every few years and the work in this section was driven by the observation that identical datasets (images from the pilot study described in chapter 3) with the same preprocessing and statistical models yielded markedly different SPMs in SPM99 and SPM2. Since in both cases preprocessing included modulation, the same level of smoothing, and the same statistical models and contrasts it seemed likely that differences in the normalisation or segmentation algorithms were behind the differences in findings. At the time of starting this work a beta version of SPM5 had been released, with a new, iterative normalisation and segmentation algorithm which was said to yield better results than the sequential application of these steps used in SPM2, and SPM2 optimised (Ashburner and Friston 2005). Therefore in order to compare the effects of the different normalisation and segmentation steps the current dataset was processed using SPM2, SPM2 optimised and SPM5, holding other options (modulation, smoothing and the statistical models) constant.

6.1.1.1 Methods

6.1.1.1.1 Subjects and assessments

Subjects were the same as those in chapter 4, and demographic data are shown in Table 4-1, page 135. Subjects had clinical, cognitive and MRI assessments as detailed in chapter 4 although only basic demographic data and the MR images were used in the work presented here.

6.1.1.1.2 VBM analysis

Good *et al.* (2001b) advocated improving the default normalisation and segmentation in SPM in two ways: firstly by making study-specific templates for both processes, which may be particularly relevant for studies of

neurodegeneration where patient neuroanatomy differs markedly from that of the healthy controls used to construct the default whole-brain templates and tissue probability maps; secondly by estimating parameters for normalisation from the tissue class of interest, rather than the whole brain (e.g. normalising grey matter to a grey matter template to "optimise" the alignment of that particular tissue class). This is easily implemented in SPM2 through a toolbox written by Christian Gaser and available at http://dbm.neuro.uni-jena.de/vbm/. With the advent of SPM5 and the new "unified" segmentation and normalisation algorithm (Ashburner and Friston 2005) "optimised" VBM as described by Good *et al.* was deemed unnecessary; improvements in the new algorithm were judged to make study-specific templates and normalisation by tissue class unnecessary.

Consequently the initial investigation compared SPM2, SPM2 "optimised" (SPM2 opt) and SPM5. The different templates used for normalisation and segmentation are summarised in Table 6-1.

Table 6-1 Summary of normalisation and segmentation templates

	Normalisation	Segmentation
SPM2	MNI-152 whole-brain	MNI-152 prior probability maps
SPM2 opt	Study-specific GM template (average of 20 controls, 10 premanifest and 10 early HD subjects, matched for age and gender)	Study-specific prior probability maps based on the same subset of subjects
SPM5	MNI-152 whole-brain	MNI-152 prior probability maps

Images from all 81 subjects in the study were segmented and normalised using the respective SPM routines and templates outlined in Table 6-1. All were then

modulated, masked to exclude non-brain (using a binary region derived from the semi-automated segmentation method described in Appendix 2) and smoothed with a kernel of 8mm FWHM. For each method the same linear regression model was used to assess differences between the groups. Voxel intensity, V, was modelled as a function of group, controlling for age, gender and TIV (an index of head size) by including them as covariates. TIV was measured as described in Appendix 2. This model is shown in equation 6.1, where the contrasts of interest are the one-tailed t-tests between the estimates of the group parameters, $\beta_1 > \beta_3$, i.e., where the controls had more volume than the early HD group.

$$V = \beta_1 \text{ Control} + \beta_2 \text{ PM} + \beta_3 \text{ HD} + \beta_4 \text{ age} + \beta_5 \text{ gender} + \beta_6 \text{ TIV} + \mu + \varepsilon \qquad (6.1)$$

(where μ is a constant, and ε is error)

The reverse contrasts were also tested, i.e. $\beta_3 > \beta_1$, looking for regions in which the early HD group had more volume than controls. Given that HD is neurodegenerative and that there are no reports of increased brain volume in symptomatic HD subjects any such findings would be most likely caused by misregistration or misclassification, rather than reflecting true anatomical differences. Statistical parametric maps were corrected for multiple comparisons using family-wise error (FWE) $p<0.05$; for the purposes of this section the exact level of correction was less important than the fact that all models were corrected to the same extent, in order to make the resulting SPMs comparable.

6.1.1.2 Results

6.1.1.2.1 Optimised vs. standard SPM

Regions in which controls had significantly more grey matter volume than the early HD group, and regions in which the early HD group had significantly more grey matter volume than controls, are shown in Figure 6-1 for each of the three SPM implementations tested.

All SPMs are centred at the maximum t value, with the exception of the HD > Control contrast for SPM5 for which the maximum t value was in the cerebellum and centring there precluded display of regions of increased volume in the cerebrum.

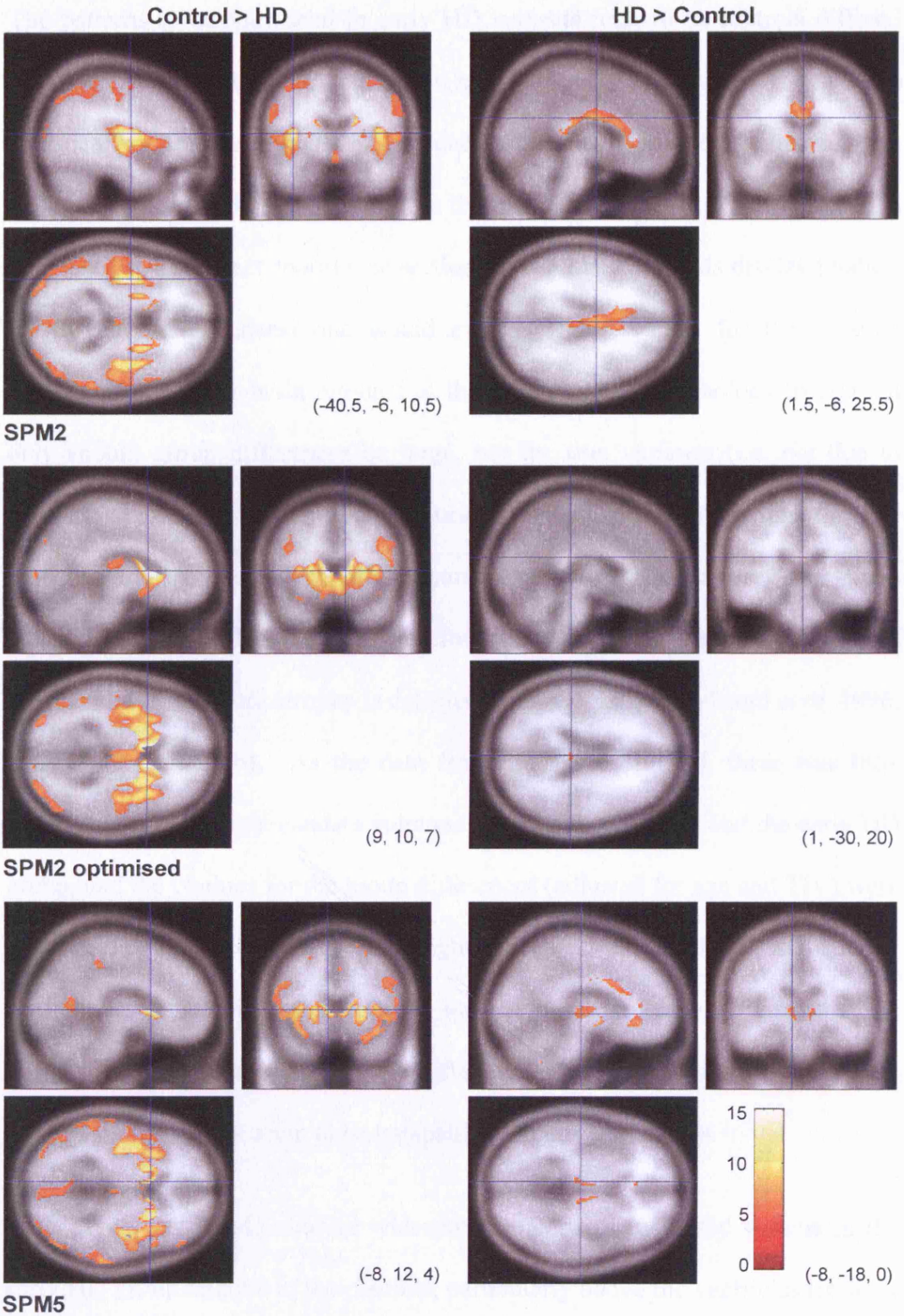


Figure 6-1 Regions of increased volume in controls relative to HD subjects (left panel) and HD subjects relative to controls (right panel) using SPM2, SPM2 optimised and SPM5.

SPMs are corrected for multiple comparisons, FWE $p<0.05$. The colour bar shows the t value. MNI coordinates are in brackets.

The patterns of atrophy seen in early HD subjects relative to controls differed between SPM versions. In particular neither SPM2 or SPM5 detected large statistically significant effects in the caudate nucleus bilaterally, compared with SPM2 optimised, although results from the previous chapter confirmed that such differences exist in this cohort (see section 5.3). Although SPMs display t values (rather than effect sizes) one would expect large t values for the striatum compared with other brain regions; as the primary site of pathology in HD not only should group differences be large, but the true variance (i.e. not due to processing errors such as misregistration or misclassification) within the HD group should be relatively small compared to that in other brain regions, given that striatal atrophy has been ongoing for many years, and that striatal volumes are reduced long before atrophy is detected in other regions (Aylward *et al.* 1996; Paulsen *et al.* 2006b). As the data from chapter 5 showed, there was little overlap in left and right caudate volumes between the controls and the early HD group, and the t values for the group differences (adjusted for age and TIV) were 15.28 (left caudate) and 17.52 (right caudate), both highly statistically significant. Although this is based on total volumes, not a voxel-wise analysis, VBM results in which there is little difference between the caudate volumes in the two groups do not seem to be compatible with these findings.

Both SPM2 and SPM5 showed widespread regions of increased volume in the early HD group relative to the controls, particularly above the ventricles (in what looks like white matter on the average template), and in regions just below the caudate bilaterally. In contrast, SPM optimised showed very little increased volume in the early HD group. Although there have been reports of increased grey matter volume in very far-from-onset premanifest HD (Paulsen *et al.*

2006b), and reports of relatively spared prefrontal cortex in early HD (Mühlau *et al.* 2007b), there are no reports of an absolute increase in grey matter after motor symptom onset.

These findings suggest that there is an element of misregistration and/or misclassification in SPM2 and SPM5, in this particular cohort, which is leading to spurious findings. The optimisation process appeared to reduce this effect. In order to investigate this further, normalised grey matter segments for two controls and two early HD subjects from SPM2 were compared (Figure 6-2).

Although the four segments are centred on the same coordinates, the cross-hairs appear to fall just underneath grey matter in the two controls, and within the grey matter of the two early HD subjects. Although this is subtle, this could explain why the early HD subjects had more grey matter than controls in this region in the SPM2 analysis. The same issues are likely to underlie the similar findings in SPM5.

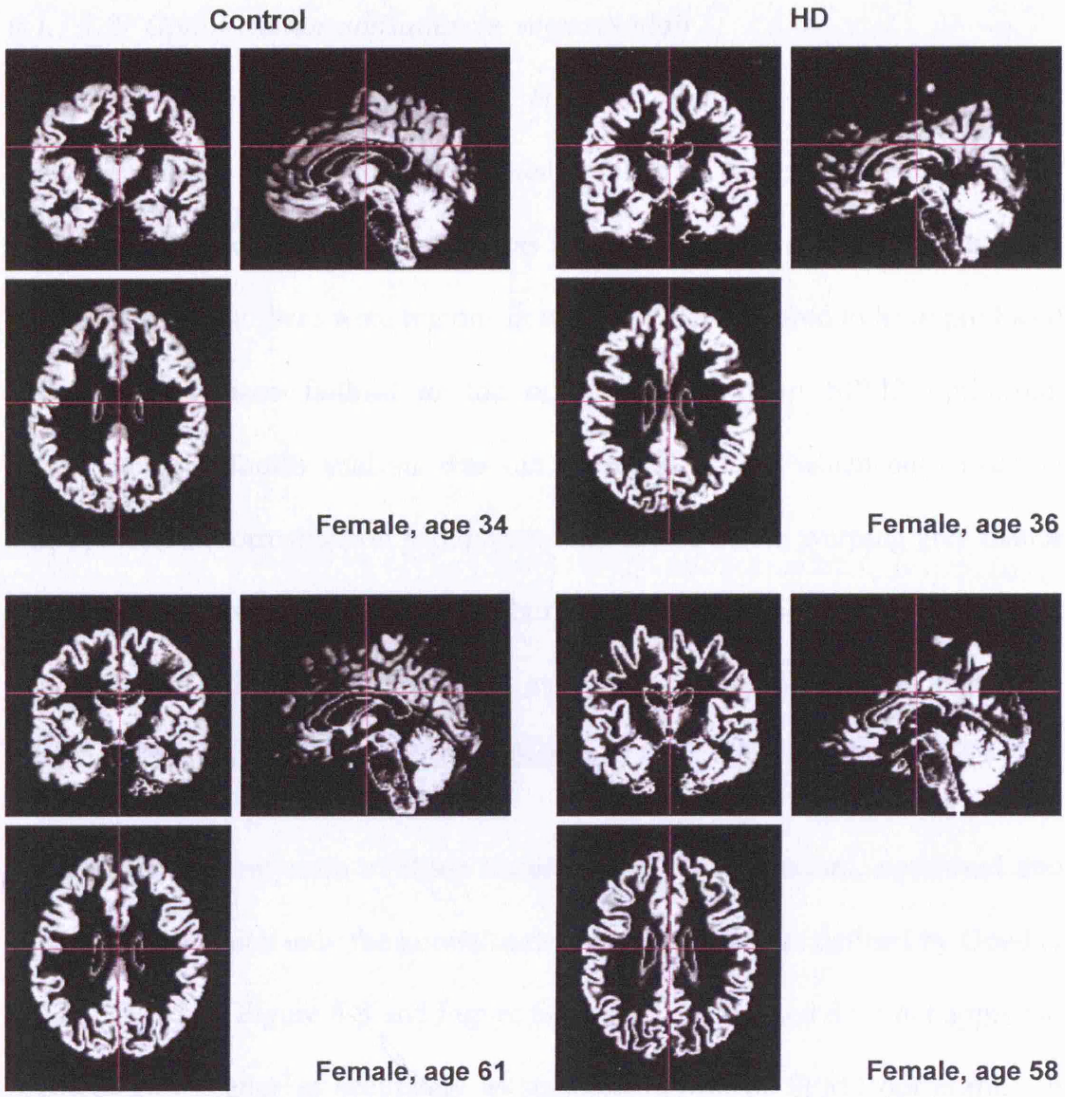


Figure 6-2 SPM2-normalised grey matter segments for two controls and two age- and gender-matched early HD subjects.

Note that whereas the cross-hairs in the controls appear just under grey matter, the equivalent voxel in the HD subjects appears within grey matter; although subtle, such differences might result in the finding of increased grey matter in HD relative to controls at this point.

6.1.1.2.2 Optimised normalisation vs. segmentation

Based on these three analyses it appeared that of the three software implementations tested, SPM2 optimised resulted in the best alignment of the subjects in this cohort. However it was also noted whilst looking at individual segmentations that there were regions in which SPM2 appeared to have produced segmentations more faithful to the original image than SPM2 optimised. Consequently a fourth analysis was run, using SPM2, in which normalisation was optimised (normalisation parameters were derived from warping grey matter segments to a grey matter template) but segmentation was not (SPM2 default templates and prior probability maps, as opposed to study-specific ones, were used for both normalisation and segmentation).

Segmentations from each of these versions of SPM2 (standard, optimised and "opt norm" in which only the normalisation was optimised as defined by Good *et al.*) are shown in Figure 6-3 and Figure 6-4. SPM2 optimised does not appear to segment grey matter as accurately as standard SPM2, or SPM2 opt norm. In addition, there are regions (notably the putamen in one subject) which SPM2 opt norm appears to segment better than either other version.

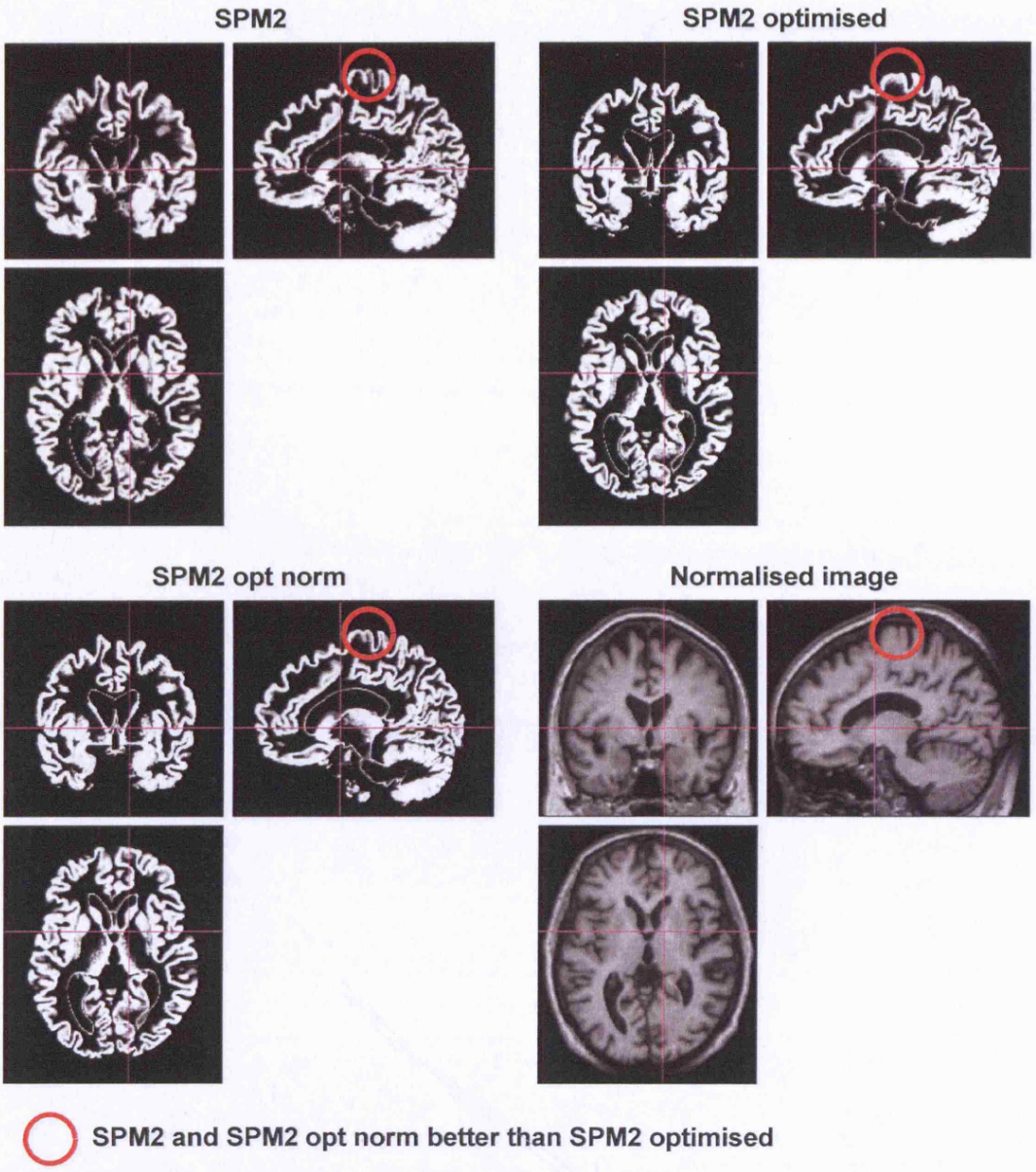


Figure 6-3 Normalised grey matter segments from SPM2, SPM2 optimised and SPM2 opt norm, displayed with a normalised image.

The red circles highlight a region in which the folds of the gyri appear to be better defined with SPM2 and SPM2 opt norm, compared with SPM2 optimised. The image is from a female HD subject, aged 64.

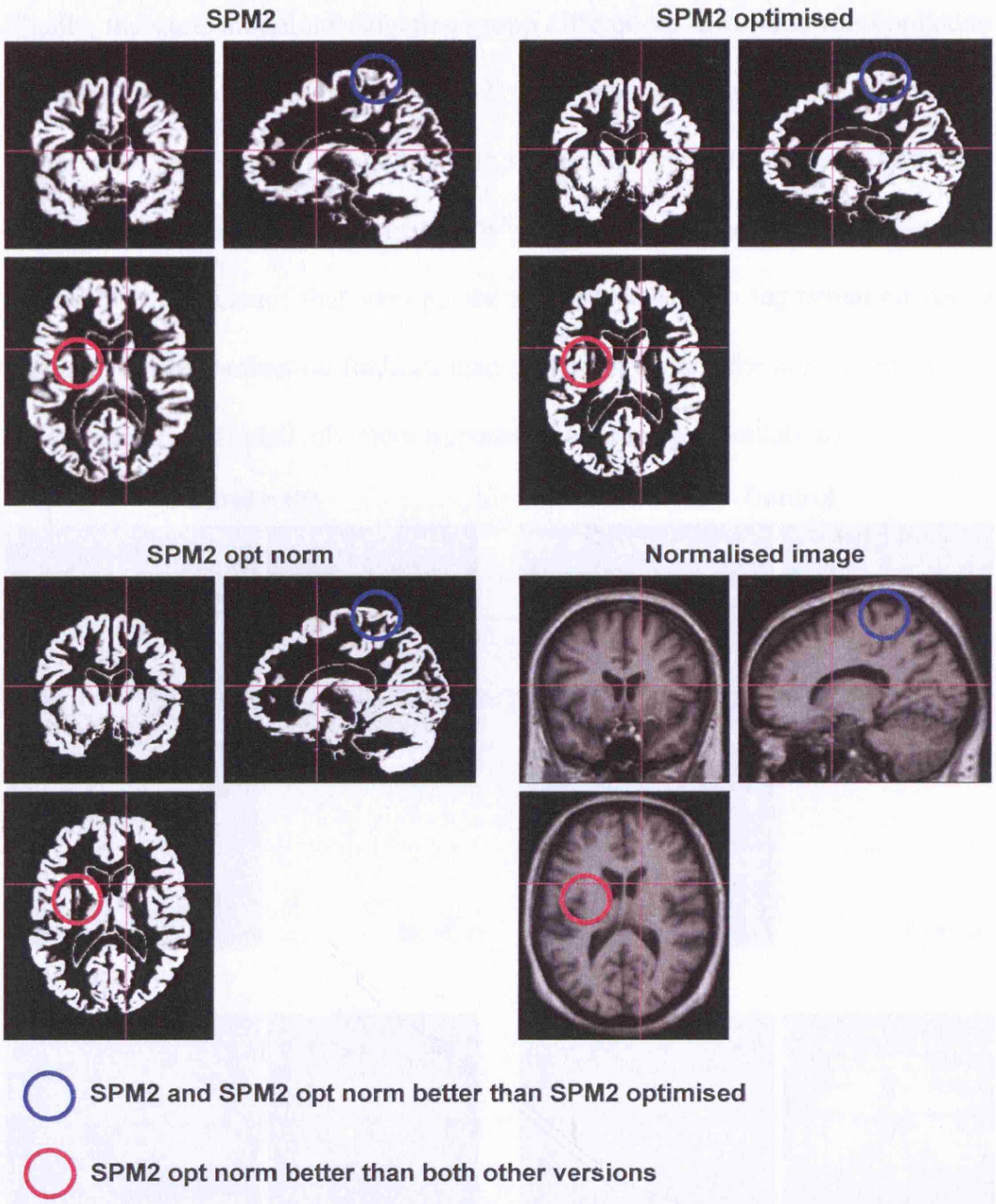


Figure 6-4 Normalised grey matter segments from SPM2, SPM2 optimised and SPM2 opt norm, displayed with a normalised image.

The blue circles highlight a region in which the folds of the gyri appear to be better defined with SPM2 and SPM2 opt norm, compared with SPM2 optimised. The pink circles show a region in which SPM2 opt norm has picked out the putamen better than SPM2 optimised, and SPM2 (which appears to miss it entirely). The image is from a female HD subject, aged 36.

Finally, the same model investigating group differences in volume was applied to the segments produced by the SPM2 opt norm procedure, and Figure 6-5 displays the resulting SPM together with that produced by SPM2 optimised. The two SPMs differ very little, particularly compared with the differences seen in Figure 6-1, suggesting that varying the template used for segmentation had a relatively smaller effect on findings than varying that used for normalisation (i.e. normalisation was relatively more important here than segmentation).

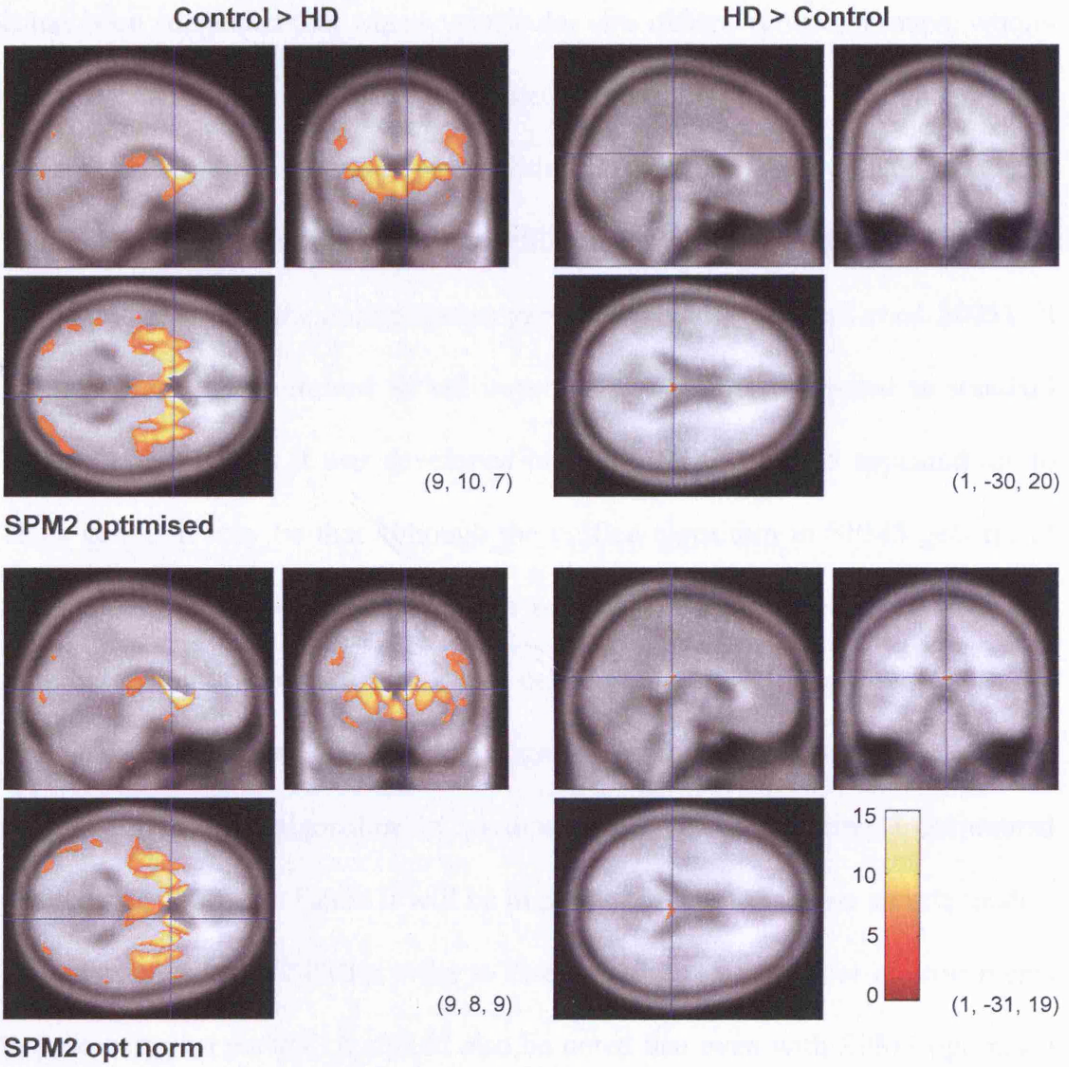


Figure 6-5 Regions of increased volume in controls relative to HD subjects (left panel) and HD subjects relative to controls (right panel) using SPM2 optimised and SPM2 opt norm.

Images are corrected for multiple comparisons, $p<0.05$. The colour bar shows the t value. MNI coordinates are in brackets.

6.1.1.3 Discussion

This section began by comparing three different implementations of SPM, and it was shown that each version produced results that differed with respect to both the extent of voxels reaching significance and the location of maxima. Notably two versions (SPM2 and SPM5) found much more extensive regions of increased volume in HD compared with SPM2 optimised, and showed fewer significant differences in the caudate bilaterally. These phenomena are likely to be related; it has been suggested that where ventricular size differs between groups, whole-brain normalisation can lead to erroneous normalisation of the surrounding tissues (Mechelli *et al.* 2005). Normalising grey matter to grey matter has been suggested as a solution for this, since this prevents differences in ventricle size from contributing to the normalisation parameters at all (Mechelli *et al.* 2005). It is interesting that optimised SPM2 improved the results compared to standard SPM2 (which is why it was developed initially) but that SPM5 appeared not to be as good. It may be that although the unified algorithm in SPM5 gets rid of some of the errors associated with the somewhat circular processing in SPM2 optimised, it still introduces some bias between the healthy and diseased brains. The problems in SPM5 have been recognised by the authors and improvements to the registration algorithm in particular are currently being incorporated (Ashburner 2007). In future it will be important to compare these simple models on newer versions of SPM in order to assess the impact of further improvements to processing on results. It should also be noted that even with SPM2 optimised and "opt norm" there was still a small region in which the HD group had increased volume relative to controls. As discussed above, there is no evidence from other modalities that manifest HD subjects show increased brain volume

relative to controls and so this is also likely to be due to imperfect normalisation, but the error is much reduced compared with the other implementations tested here.

In the current study normalising the images to a study-specific grey matter template reduced misalignment (judged by looking at regions in which the HD group were shown to have more volume than controls). However this improvement remained when the images were normalised to a non-study-specific (i.e. healthy average) grey matter template, suggesting that in this cohort the main improvement was achieved by maximising the alignment of the tissue class, rather than by making the template more representative of the cohort. A similar finding was reported by Keller *et al.* (2004) who also concluded that the better alignment provided by grey matter to grey matter normalisation had far more influence on results than the choice of segmentation template. Whilst it is true that in studies of neurodegeneration, use of a healthy whole-brain template might lead to systematic processing bias between controls and patients, it has also been suggested that to be representative a study-specific template would need to consist of a minimum of ~100 images (J. Ashburner, personal communication). Most studies cannot provide such subject numbers, and may therefore be introducing another source of error by using study-specific templates based on too few subjects. There were subtle differences between the segmentations of all three versions of SPM2, and in particular it was noted that the putamen for one subject was segmented better using SPM2 opt norm, rather than SPM2 opt. Assuming that images are well-normalised to standard space after GM to GM normalisation, it seems reasonable that segmentation with a prior based on >100 subjects might be better than with one based on 40.

As a result of this work the SPM2 opt norm procedure was used for all the subsequent VBM work presented in this thesis, including in the following sections. The important recommendation from this work is to assess reverse contrasts in SPM comparisons and to recognise the degree of dependence of the findings upon the analysis method used, including the specific version of the algorithms used for pre- and post-processing, and to report these.

6.1.2 Effect of head size and gender on brain volume

Apart from the effects of pathology, total brain volume in "healthy" subjects is known to vary with both head size (Acer *et al.* 2007) and gender (Good *et al.* 2001a) and it has been shown that adjusting whole-brain volume for TIV eliminates differences due to gender (Whitwell *et al.* 2001). In Whitwell *et al.* the adjustment was simply to divide brain volume by TIV. It is common for volumetric studies to include an adjustment for some index of head size in order to ensure that differences in head size are not influencing findings (e.g., Kassubek *et al.* 2004a; Paulsen *et al.* 2006b). However a number of VBM studies in HD and other areas do not adjust for head size.

After normalisation in VBM, brains (or grey matter segments) have been warped to the same global space, and at this point differences due to head size (or the size of the brain initially) are lost. If a voxel from a large brain was compressed during normalisation, there would then be nothing to suggest a difference between that voxel and one from a smaller brain, if they both had the same probability of being grey matter. However after the voxel value has been multiplied by the Jacobians from normalisation, intensity at that voxel then represents volume at that point (this is known as modulation). After modulation

the voxel intensity is increased, to represent the fact that it has had to be relatively compressed to match the template, and in this way differences between volume, which were previously encoded spatially, are preserved in the voxel intensities. Therefore, with modulated data, it would seem necessary to adjust for a measure of head size.

The work in this section investigated the effect of TIV and gender on volume in a VBM analysis, and also looked at whether adjusting for TIV eliminated effects of gender on volume.

6.1.2.1 VBM analysis

This section used the images of the 20 healthy control subjects. Images were preprocessed as described above (using the SPM2 opt norm procedure) and were entered into the following models to look at the effect of head size and gender on brain volume.

$$V = \beta_1 \text{ TIV} + \mu + \varepsilon \quad (6.2)$$

$$V = \beta_1 \text{ gender} + \mu + \varepsilon \quad (6.3)$$

$$V = \beta_1 \text{ TIV} + \beta_2 \text{ gender} + \mu + \varepsilon \quad (6.4)$$

In the first two models, the contrast $\beta_1 > 0$ investigates regions in which the slope of the relationship between volume and TIV (or gender) is significantly greater than zero (i.e. where larger TIV (or being male) is associated with larger brain volume). The third model (6.4) allows investigation of the effect of each variable on volume, having adjusted for the effects of the other, by examining $\beta_1 > 0$ and $\beta_2 > 0$. For all models the reverse contrast was also investigated, i.e.

looking for regions in which a volume decrease was associated with increasing head size or male gender.

6.1.2.2 Results

Regions in which grey matter volume increased with TIV, male gender and TIV having adjusted for gender are shown in Figure 6-6. Larger TIV was associated with increased grey matter volume in the cortex and subcortical structures. Volume also tended to be increased in men compared to women, although the effects were less widespread. When gender was adjusted for, there was still an effect of TIV on volume although it was less extensive. When TIV was adjusted for there was no longer an effect of gender on volume.

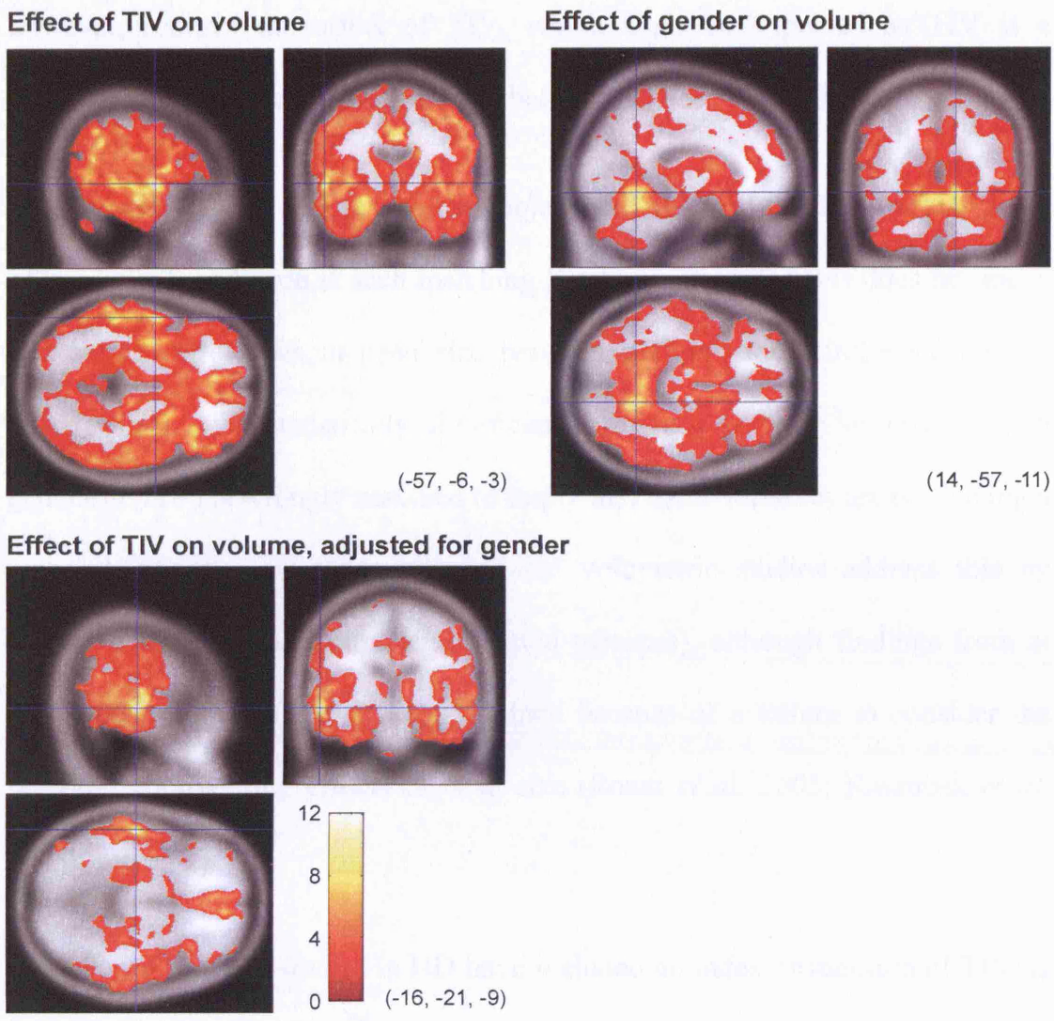


Figure 6-6 Regions in which grey matter volume increased with TIV, gender (male), and with TIV after adjusting for gender.

There were no regions in which volume increased with maleness, after adjusting for TIV. SPMs are corrected for multiple comparisons, FDR $q<0.05$. The colour bar shows the t value. MNI coordinates are in brackets.

6.1.2.3 Discussion

In the 20 controls in this cohort there was a clear effect of both TIV and gender on grey matter volume. TIV is known to correlate with gender, and adjusting for TIV removed the effect of gender, which agrees with previous findings using whole-brain volumes (Whitwell *et al.* 2001). Adjusting for gender alone did not,

however, remove the effect of TIV, which is to be expected as TIV is a continuous variable and captures more between-subject variability than gender.

Few imaging studies report whether subjects are matched for TIV (or any index of head size), and even if such matching has been attempted this does not mean that subtle differences in head size between groups cannot influence results. Often a lack of a statistically significant difference between groups (e.g., in gender or TIV) is wrongly assumed to imply that these variables are not having a material influence on the results. Most volumetric studies address this by adjusting for TIV (or a similar correlated measure), although findings from at least one study in HD have been doubted because of a failure to consider the potential confounding effects of head size (Rosas *et al.* 2003; Kassubek *et al.* 2004b).

However, few VBM studies in HD have included an index or measure of TIV as a covariate although many adjust for total grey matter volume. In healthy subjects total grey matter volume is likely to correlate with TIV (although will decrease with age) and so assuming age is adjusted for, adjusting for total grey matter volume approximates an adjustment for TIV and allows investigation of differences in grey matter volume that are not caused by differences in overall head size. However in subjects with a neurodegenerative disease total grey matter volume will almost certainly decrease with duration or severity of the disease, and hence adjusting for it is likely to mask some disease-related effects. At an extreme level, if degeneration proceeded uniformly throughout the brain then a comparison between healthy controls and patients that was adjusted for total grey matter volume would find no evidence of group differences.

Adjusting for total grey matter volume does allow investigation of the relative loss or preservation of regions, compared with the amount of global loss (Good *et al.* 2001b; Mechelli *et al.* 2005). This is an interesting question in itself, but needs careful interpretation. Some studies seem to equate adjustment for total grey matter volume with adjustment for head size, but in studies of neurodegeneration in particular, adjusting for the former will get rid of some disease-related effects, whereas adjusting for the latter will not.

It is important to note that the adjustment for TIV in the controls presented here also corrects for differences in brain volume that are due to gender. However if there is evidence that a disease characteristic also varies with gender then it may still be worth including gender as a covariate, either to adjust for these effects or to investigate them.

6.1.3 Measuring TIV

The previous section highlighted the importance of adjusting brain volumes for head size. In this thesis TIV is used as an index of head size and is measured according to the protocol in Appendix 2. Manual outlining is generally assumed to be the "gold standard" measure of volume, but it is also labour-intensive and time-consuming, taking time to train segmentors and needing about 30 minutes per TIV. In the protocol used here every 10^{th} slice was segmented, and linear interpolation was used to estimate the total TIV, which speeds up the segmentation without resulting in a significant decrease in accuracy (Whitwell *et al.* 2001).

However SPM also generates a measure of TIV: the sum of the GM, WM, and CSF segments that are produced during preprocessing. These segments have

been used previously to estimate total intracranial volume in HD (Kassubek *et al.* 2004a; Mühlau *et al.* 2007b) although some doubt has been expressed as to the accuracy of the CSF segmentations in particular (J. Ashburner, personal communication). Given that manually-derived TIVs and SPM-derived GM, WM and CSF segments were available for all the subjects in this study the purpose of this section was to compare the two estimates of head size. As for the purposes of the work in this thesis TIV was used as a nuisance covariate, rather than a variable of interest, it would not be vital for the two estimates to agree perfectly, but it would be important for them to be highly correlated, and for this relationship to be consistent in both healthy controls and HD subjects. If this were found it might mean that in the future TIV could be estimated using the SPM segments, thus saving a large amount of operator time.

6.1.3.1 Methods

TIV was measured manually according the protocol described in Appendix 2. TIV was also estimated as the sum of the GM, WM and CSF volumes generated by the VBM preprocessing. This was done for all 81 subjects for whom baseline data were collected (see chapter 4).

6.1.3.1.1 Statistical analysis

The extent of association between the two measures was assessed using a correlation coefficient. Pitman's test was used to compare the variance of each measure, both for the cohort as a whole and within each subject group. Agreement between the two measures was assessed through calculation of the mean and standard deviation of the difference between measures, again for the cohort as a whole and for each subject group (Bland and Altman 1986). Linear

regression models were used to investigate whether the slope of the relationship between the two measures differed between groups (an interaction between slope and group) and/or whether the intercept differed between groups (as would occur if bias between the measures was not equal across the three groups).

6.1.3.2 Results

Mean (SD) TIVs from each method are presented in Table 6-2. The two measures were highly correlated for the whole cohort ($r=0.94$, $p<0.0001$). There was no evidence that the variance differed between methods, either for the whole cohort or for any of the subject groups (all $p>0.1$). SPM estimates were, on average, larger than the manual estimate for the entire cohort (bias (mean difference) = 34.7ml, 95% CI 22.2, 47.2ml, $p<0.0001$) and this was driven by the early HD group (bias = 58.4ml, 95% CI 39.2, 77.6ml, $p<0.0001$); bias in the control and premanifest groups was smaller and not statistically significant (all $p>0.05$).

Table 6-2 Mean (SD) TIVs for the whole cohort and each group, with differences

TIV (ml)	All (N=81)	Control (N=20)	Premanifest (N=21)	Early HD (N=40)
Manual	1448 (162)	1434 (164)	1469 (182)	1445 (153)
SPM	1483 (157)	1452 (155)	1474 (167)	1503 (155)
Difference (bias)	34.7 (56.4)	18.2 (40.5)	5.3 (42.8)	58.4 (60.0)
95% CI	22.2, 47.2	-0.8, 37.2	-14.2, 24.8	39.2, 77.6

Positive difference indicates higher estimates using SPM relative to manual measures.

There was no evidence that the slope of the relationship between manual- and SPM-derived TIVs differed between groups. Constraining the slope to be the same in each group, the intercept was greater in the early HD group compared with controls ($p=0.006$) and the premanifest group ($p<0.0001$), reflecting the fact that the bias in this group was significantly greater than in the other two.

Therefore, compared with manual measures, SPM overestimated TIV, and overestimated it significantly more in the early HD group than in the other two groups.

6.1.3.3 Discussion

This work demonstrates that the TIV measure derived from the sum of the SPM GM, WM and CSF segments is significantly larger than the manual measure. More importantly the amount by which SPM overestimates TIV differs significantly between healthy controls and early HD subjects.

As discussed in the introduction the fact that the SPM measure overestimates *per se* would not preclude it from being used as a nuisance covariate in place of the manual measure, particularly as the two were highly correlated. The overestimation may be because the CSF segments from SPM often spill outside the dura, whereas the manual measure excludes CSF outside dura. However it is worrying that the SPM measure overestimates TIV more in the early HD group (the group with the most atrophy) relative to the controls and premanifest subjects, as this could lead to a systematic bias in group differences when volume was adjusted for TIV. This may be because in the early HD group there is a greater discrepancy between the normalised brain and the tissue prior probability maps (because of increased atrophy) and this leads to relatively greater

misclassification of non-brain and dura as tissue. In this cohort, this would affect the volume of the segmentations, although would not affect the final analysis as segments are masked to exclude non-brain tissue prior to analysis. However this bias has only been demonstrated on the current dataset, and further work needs to be done to investigate whether results are similar in cohorts with different amounts and patterns of atrophy.

This interpretation assumes that the manual measure is the more accurate representation of TIV. In fact, a similar manual measure of TIV on CT images was shown to underestimate true TIV (measured directly from dry skulls) (Sahin *et al.* 2007). However in the absence of real skull measurements on the cohort presented here manual measures have to be taken as the "gold standard" estimate of TIV, and in addition there is no *a priori* reason why manual segmentation (which is done blinded to gene status) should have systematically underestimated TIV only in the early HD group, which would be the alternative explanation for the pattern of results seen here.

Overall this work demonstrated that for this cohort SPM-derived TIVs were not an adequate alternative to manually-derived ones. This may have implications for other users of SPM-derived TIV measures but these findings need to be replicated in other cohorts.

6.2 GENERAL DISCUSSION AND REVIEW

6.2.1 Technical considerations

The work in this chapter has demonstrated a number of important methodological points about VBM analyses. Firstly the SPM version can affect the final results, and even within a version, it is worth investigating the impact of

different normalisation and segmentation templates in order to maximise the accuracy of these steps for the cohort under investigation. A version of SPM2 in which normalisation was specific to the tissue class under investigation, and standard (rather than study-specific) templates were used for segmentation, was preferred to both standard SPM2 and SPM5, and SPM2 optimised. As future versions of SPM are released it will not be possible to rule out software differences as causes for differences between studies, but it would be useful for investigators to study the effects of different preprocessing options and in particular to try and maximise normalisation and segmentation accuracy for their cohort.

Secondly, although variables such as head size and gender are often adjusted for in volumetric analyses, they are less often considered as possible confounds in VBM analyses. However when modulated data were used there was a clear effect of head size and gender on the controls in this cohort. In contrast, it is common for VBM studies to adjust for total grey matter volume. Although this allows investigation of regionally-specific patterns of atrophy, and relative decrease or preservation of tissue, this is not an alternative to correcting for differences due to head size where neurodegenerative diseases are under investigation, as total grey matter volume will vary with disease severity and duration.

Finally, more work needs to be done to assess the validity of using SPM segments as an estimate of TIV, as work on this cohort suggested that SPM may introduce systematic differences between groups in its estimate of TIV.

6.2.2 Implication for differences between findings in HD

Within the relatively few published VBM studies in HD there are differences at almost every step, from the software version used, through the processing steps and statistical models, to correction for the resulting SPM. Table 6-3 summarises some of the processing methods and levels of correction used in HD VBM studies, including studies that reported group differences, and those which investigated neural correlates of variables, since both types of analyses are used in this thesis. These differences make it hard to interpret the various findings, and may mean that results do not generalise to the population as a whole.

Table 6-3 Summary of processing methods employed by other groups

Study	SPM version	Normalisation	Segmentation	Mod.	Smoothing FWHM, mm	Correction
Thieben *et al.* (2002)	99	Study-specific GM template, patients and controls	Unspecified	Yes	10	SPM uncorrected, $p<0.005$; reported results mostly small-volume corrected
Ho *et al.* (2004)	99	Study-specific GM template, all controls only	Study-specific GM template, all controls only	Yes	12	SPM & reported results uncorrected, $p<0.0001$, cluster 10 voxels
Kassubek *et al.* (2004c)	99	Study-specific template, whole-brain or GM unspecified; subjects unspecified	Unspecified	No	6	SPM & reported results FWE $p<0.001$, clusters 54 voxels
Kassubek *et al.* (2005)	99	Study-specific template, 50:50 patients:controls; whole-brain or GM unspecified	Unspecified	Yes	6	SPM & reported results FWE $p<0.001$

Study	SPM version	Normalisation	Segmentation	Mod.	Smoothing FWHM, mm	Correction
Kipps *et al.* (2005)	2b & 99	Study-specific GM template, 50:50 patients:controls	Standard implied	Yes	12	SPM uncorrected, $p<0.005$; reported results small-volume corrected (level unspecified)
Peinemann *et al.* (2005)	99	Study-specific template, whole-brain or GM only unspecified; subjects unspecified	Unspecified	No	6	SPM & reported results FWE $p<0.05$
Douaud *et al.* (2006)	2	Study-specific GM template, 50:50 patients:controls, from original & symmetric images	Study-specific GM template, 50:50 patients:controls, from original & symmetric images	Yes	8	SPM & reported results FDR $q<0.01$
Barrios *et al.* (2007)	Not specified	Standard whole-brain template	Not specified	No	4	SPM & reported results uncorrected, $p<0.01$, clusters > $10mm^3$

Study	SPM version	Normalisation	Segmentation	Mod.	Smoothing FWHM, mm	Correction
Gavazzi *et al.* (2007)	2	Study-specific GM template, subjects unspecified	Study-specific GM template, subjects unspecified	Yes	10	SPM and reported results corrected, $p<0.01$, (type unspecified)
Jech *et al.* (2007)	2	Study-specific GM template, all patients; no controls in study	Study-specific GM template, all patients; no controls in study	Yes	10	SPM uncorrected $p<0.001$; reported results uncorrected in striatum or rolandic area, $p<0.001$, elsewhere, FDR $q<0.05$
Kipps *et al.* (2007)	2	Study-specific template, patients and controls, exact makeup unspecified	Not specified	Yes	8	Uncorrected, $p<0.05$
Mühlau *et al.* (2007b)	2	Study-specific prior probability maps, subjects and whether used for normalisation as well as segmentation unspecified		Yes	8	SPM & reported results FWE $p<0.05$, clusters $p<0.05$

Study	SPM version	Normalisation	Segmentation	Mod.	Smoothing FWHM, mm	Correction
Mühlau *et al.* (2007a)	2	Study-specific prior probability maps, subjects and whether used for normalisation as well as segmentation unspecified		Yes	8	SPM & reported results FWE $p<0.05$, extent $p<0.05$, clusters $p<0.001$
Ruocco *et al.* (2007)	2	Study-specific GM template, healthy volunteers otherwise unused in study	Study-specific GM template, healthy volunteers otherwise unused in study	Yes	10	SPM & reported results FDR $q<0.05$
Wolf *et al.* (2007)	2	Study-specific whole-brain template, 50:50 patients:controls	Study-specific GM templates, 50:50 patients:controls	Yes	8	SPM not shown; reported results FWE $p<0.001$
Wolf et al. (2008)	2	Study-specific whole-brain template, 50:50 patients:controls	Study-specific GM templates, 50:50 patients:controls	Yes	8	SPM & reported results FWE $p<0.001$

Mod. = modulation

Firstly, SPM itself changes as newer versions are released, as has been demonstrated in the first part of this chapter. Whilst this is unavoidable these upgrades can incorporate changes to the segmentation algorithms, and the statistics, which can mean that the same dataset analysed in different versions will not necessarily yield identical SPMs. This needs to be taken into account when comparing results from earlier studies.

Secondly, there are obvious differences in the preprocessing. Although SPM is automated in that there is very little manual processing of images, there are a number of options that can be set by the user. As discussed above, the original paper to advocate an "optimised" normalisation and segmentation for VBM (Good *et al.* 2001b) described normalising grey matter to grey matter to get better alignment, and also using a study-specific grey matter template for both this normalisation and the subsequent segmentation. However since then a range of "optimisation" methods have been reported: some studies just optimise the normalisation but do not refer to segmentation, some use a study-specific template but for the whole brain rather than different tissue classes, and some introduce other new steps (compare e.g., Thieben *et al.* 2002; Peinemann *et al.* 2005; Douaud *et al.* 2006). Since misalignment or bad segmentation could introduce false group differences, the normalisation and segmentation processes are other factors which need to be considered when comparing results from different groups. At the very least interpretation would be aided by the knowledge that preparatory work had tried to overcome (population-specific) problems with either step (e.g., Douaud *et al.* 2006).

Another important issue is that of modulation. Normalisation is supposed to correct for global differences in head position and structure (e.g. line up the left superior temporal gyri on all subjects) but not local differences due to atrophy. However in practice it is likely that normalisation will result in some atrophy being lost. To correct for this, the modulation step multiplies the voxel intensity by the Jacobian from the normalisation process (i.e. by an index of by how much that voxel was stretched or contracted) so that intensity is a more accurate representation of volume. With modulated data one is testing for ". . . regional differences in the absolute amount (volume) of grey matter . . ." (Good *et al.* 2001b, p24) whereas with unmodulated data one is looking at ". . . differences in concentration of grey matter (per unit volume in native space)." (Ashburner and Friston 2000; Good *et al.* 2001b, p24). Three of the 16 HD studies that have used VBM do not mention modulation (Kassubek *et al.* 2004c; Peinemann *et al.* 2005; Barrios *et al.* 2007) and hence the SPMs from these studies may not be showing the same sort of data as the others. There are clear differences between findings with modulated and unmodulated data, and incorporating a modulation step is the preferred way of ensuring inter-subject alignment without losing inter-group differences in morphology (Keller *et al.* 2004).

A final preprocessing option is the smoothing kernel. Smoothing is required in order to render the data more normally distributed and to correct for some error in the registration process (Ashburner and Friston 2000). Ashburner and Friston (2000) state that "Whenever possible, the size of the smoothing kernel should be comparable to the size of the expected regional differences between the groups of brains." (Ashburner and Friston 2000, p807). A kernel of 8mm was used in the current work because it was shown to give a better representation of simulated

atrophy than smaller or larger kernels (R. Scahill, PhD Thesis, University of London). The larger the kernel, the wider the region over which the signal is averaged, which tends to yield more extensive areas of significant differences but make precise anatomical localisation harder. This again means that differences between studies in which different kernels have been used might not reflect true differences in the cohorts being studied.

After preprocessing a common analysis is to regress intensity at each voxel against a group variable, including other variables in the model as necessary in order to control for their effects (e.g. age, TIV). The test of interest is the t-test between the parameter estimates for the group variable. There is huge variation between research groups in the covariates they have included in this model: some (as in this study) include age, gender, and TIV since these are factors that can affect brain volume independently of disease, but some do not. The different effects of adjusting for either head size or total grey matter volume in a disease population have been discussed earlier in section 6.1.2. Another common analysis is to use simple regression models to look at the association between a variable of interest and brain volume. Whilst some groups model this as a regression, others choose to compare the outcome of two separate contrasts, e.g. high CAG repeat length vs. controls, and low CAG repeat length vs. controls. Whilst this can yield interesting SPMs, visual comparison of the two resulting statistical maps does not constitute a valid statistical comparison in itself.

Finally, very different pictures can be obtained by varying the level and type of correction used. One of the benefits of VBM is the fact that it looks at the whole brain in an unbiased way (as has been done in this chapter) but in doing so many

thousands of statistical tests are performed at once. At a standard alpha level of 0.05, approximately 5000 voxels in an image of 100000 voxels would be expected to be false positives. One solution to this is to control the family-wise error (FWE) rate, by thresholding the statistic at a level so as to control the probability of there being at least one false positive voxel in the entire SPM, but this can lack power and hence omit many true positives (Genovese *et al.* 2002). Recently a new method has been described in which the proportion of false positives in significant voxels is controlled (the False Discovery Rate (FDR), Benjamini and Hochberg 1995). This might be particularly useful for the multiple testing problem in neuroimaging, as it is more powerful than controlling the FWE and the thresholds are determined directly from the data (rather than from the test statistic) making comparison between studies simpler (Genovese *et al.* 2002).

As can be seen from Table 6-3, alpha levels in HD VBM studies range from the conservative 0.001 (controlling the family-wise error rate) to the more exploratory 0.005 (without correction for multiple comparisons). Uncorrected results are often published where group numbers or effect sizes are small, and hence would not survive FWE correction, although as discussed above, this is likely to result in a large number of false positives. Conversely extremely stringent control of the FWE rate is likely to lead to under-reporting of true effects. Overall, the wide variation in the type and level of correction makes it difficult to compare results between studies. Recently sets of guidelines have been published to encourage more uniformity in the reporting of functional and structural imaging studies and these suggest that the multiple testing problem should be addressed clearly, and that unthresholded statistical or effect maps

should also be presented in order to aid inter-study comparison (Poldrack *et al.* 2007; Ridgway *et al.* 2008). Showing unthresholded effect or statistical maps would also make neuroimaging data more similar to those presented in other fields, where typically both significant and non-significant effects are reported, together with p values or confidence intervals. In future, both use of the FDR (which is easily interpretable across studies) and presentation of effect maps across the whole search volume, are likely to help emphasise similarities, rather than differences, between studies.

A related issue is that some studies have also opted to threshold clusters of results (for example, excluding groups of voxels that are smaller than a certain size or fail to reach significance at a certain level), although this statistic is thought to be invalid in VBM (Good *et al.* 2001b; Mechelli *et al.* 2005) as it requires uniform smoothness throughout the image. Since smoothness is variable in VBM, larger clusters are likely to be found by chance where data are more smooth, and hence this approach risks missing valid differences that fail to meet the specified cluster threshold.

6.3 CONCLUSION

One of the aims of the work in this chapter was to investigate the effect of different software versions of SPM on the HD dataset. Choice of normalisation template (whole-brain or specific to a tissue type) appeared to have the largest effect on findings, whilst choice of segmentation template had a much less obvious effect. SPM2, using a grey matter (but not study-specific) template for normalisation, resulted in the best alignment of images (judged by looking at regions in which the early HD subjects were shown to have greater volume than

controls) and marginally better segmentations (judged by visual comparison of segmentations) than the other two implementations tested. Consequently this set of VBM processing steps was used for the VBM work presented in the remainder of this thesis.

Finally, it is clear that any two VBM papers rarely use exactly the same methods and parameters which can mean that results are not directly comparable. More importantly, seemingly conflicting results may be due to methods rather than a false finding or inherent differences in the population studied. This problem is not restricted to VBM, although the methodological variations in the 16 studies in Table 6-3 illustrate the difficulties well. Recently there has been a call for work to focus on validating methods and reproducing and confirming earlier findings in order to reduce inconsistencies in the imaging literature (Nopoulos *et al.* 2007). A better consensus on some of the options highlighted in this chapter might aid interpretation and generalisation of results, and will be needed if robust conclusions are to be drawn about brain atrophy in HD.

7 PATTERNS OF ATROPHY AND THE RELATIONSHIP BETWEEN CAG REPEAT LENGTH AND ATROPHY IN HUNTINGTON'S DISEASE: A VOXEL-BASED MORPHOMETRIC STUDY

7.1 INTRODUCTION

As discussed in section 1.2.2, although it is clear that extra-striatal atrophy is present even early in HD, findings differ as to the exact location and extent of this atrophy. There are similarly disparate findings on the relationship between CAG repeat length and volume loss, using both region of interest volumetry and VBM techniques.

The work and technical review in the previous chapter demonstrated how, in the field of VBM, methodological differences can influence results, and also showed that few studies in HD have used identical preprocessing and analysis parameters, suggesting that at least some of the discrepancies between studies can be attributed to these differences. This makes it hard to judge the veracity of effects which may or may not be specific to HD. The work in the previous chapter also sought to maximise the accuracy of the normalisation and segmentation algorithms for the HD dataset and to demonstrate the effects of some confounding variables. VBM processing in this and subsequent chapters follows the "opt norm" procedure outlined in chapter 6.

The work in this chapter was undertaken to address two questions in particular: 1) using a strict level of whole-brain correction in a large cohort of subjects at different stages of HD, could atrophy be detected and where; and 2) what effect does CAG repeat length have on brain volume?

7.2 METHODS

7.2.1 Subjects

Subjects were the same 81 as those reported in chapter 4, see section 4.2.1 for details.

7.2.2 Assessments

Subjects underwent the same assessments described in chapter 4, sections 4.2.2.1, 4.2.2.2, and 4.2.2.3. That is, they underwent clinical and cognitive assessments and had 1.5T MRI scans.

7.2.3 VBM analysis

Images were preprocessed following the steps described in section 2.5.5.

7.2.3.1 Group differences

Linear regression models were used to examine differences in grey matter volume between groups. Voxel intensity, V, was modelled as a function of group, controlling for age, gender and TIV (an index of head size) by including them as covariates. TIV was measured as described in Appendix 2. Although TIV can be calculated as the sum of the GM, WM and CSF segments derived from the preprocessing stages of SPM, measurement error is likely to be increased relative to manually-delineated measures (see section 6.1.3). This model is shown in equation 7.1, with the contrasts of interest being the one-tailed t-tests between the estimates of the group parameters, i.e. $\beta_1 > \beta_2$, or $\beta_1 > \beta_3$ etc.

$$V=\beta_1 \text{ Control} + \beta_2 \text{ PM} + \beta_3 \text{ HD} + \beta_4 \text{ age} + \beta_5 \text{ gender} + \beta_6 \text{ TIV} + \mu + \varepsilon \qquad (7.1)$$

(where μ is a constant, and ε is error).

The analysis was repeated with global GM as an additional covariate to look for areas in which there was relative sparing or loss of tissue having controlled for the global effect of the disease (equation 7.2).

$$V=\beta_1 \text{ Control} + \beta_2 \text{ PM} + \beta_3 \text{ HD} + \beta_4 \text{ age} + \beta_5 \text{ gender} + \beta_6 \text{ TIV} + \beta_7 \text{ global} + \mu + \varepsilon \quad (7.2)$$

7.2.4 Atrophy and CAG repeat length

All 61 subjects with the HD gene expansion (premanifest and early HD) were entered into a simple regression model to look for associations between volume and CAG repeat length, having controlled for the effects of age, gender and TIV (equation 7.3). The contrast of interest is $\beta_1 < 0$, i.e. regions in which there is a significant negative association between volume and CAG repeat length.

$$V=\beta_1 \text{ CAG} + \beta_2 \text{ age} + \beta_3 \text{ gender} + \beta_4 \text{ TIV} + \mu + \varepsilon \quad (7.3)$$

The resulting SPMs were masked as described in section 2.5.5 and thresholded at $p<0.05$ using family-wise error (FWE) correction.

7.3 RESULTS

As subjects were the same as those reported in chapter 4, demographic data can be found in Table 4-1, page 135.

7.3.1 Early HD vs. Controls

Early manifest HD subjects had significantly less grey matter than controls (after controlling for age, gender and TIV) in the caudate, putamen and insula bilaterally, as well as the anterior cingulate and thalamus (Table 7-1, Figure 7-1A). There was also evidence of more widespread cortical atrophy, including

the medial frontal gyrus bilaterally, the left inferior frontal gyrus, the postcentral gyrus bilaterally, left parietal cortex as well as the right angular gyrus, and bilateral regions in the temporal and occipital lobes (Figure 7-1B). There was a small region of increased grey matter in the HD group relative to controls (shown in the bottom right panel of Figure 6-5) which is likely to reflect imperfect normalisation (see section 6.1.1.3 for further discussion of this).

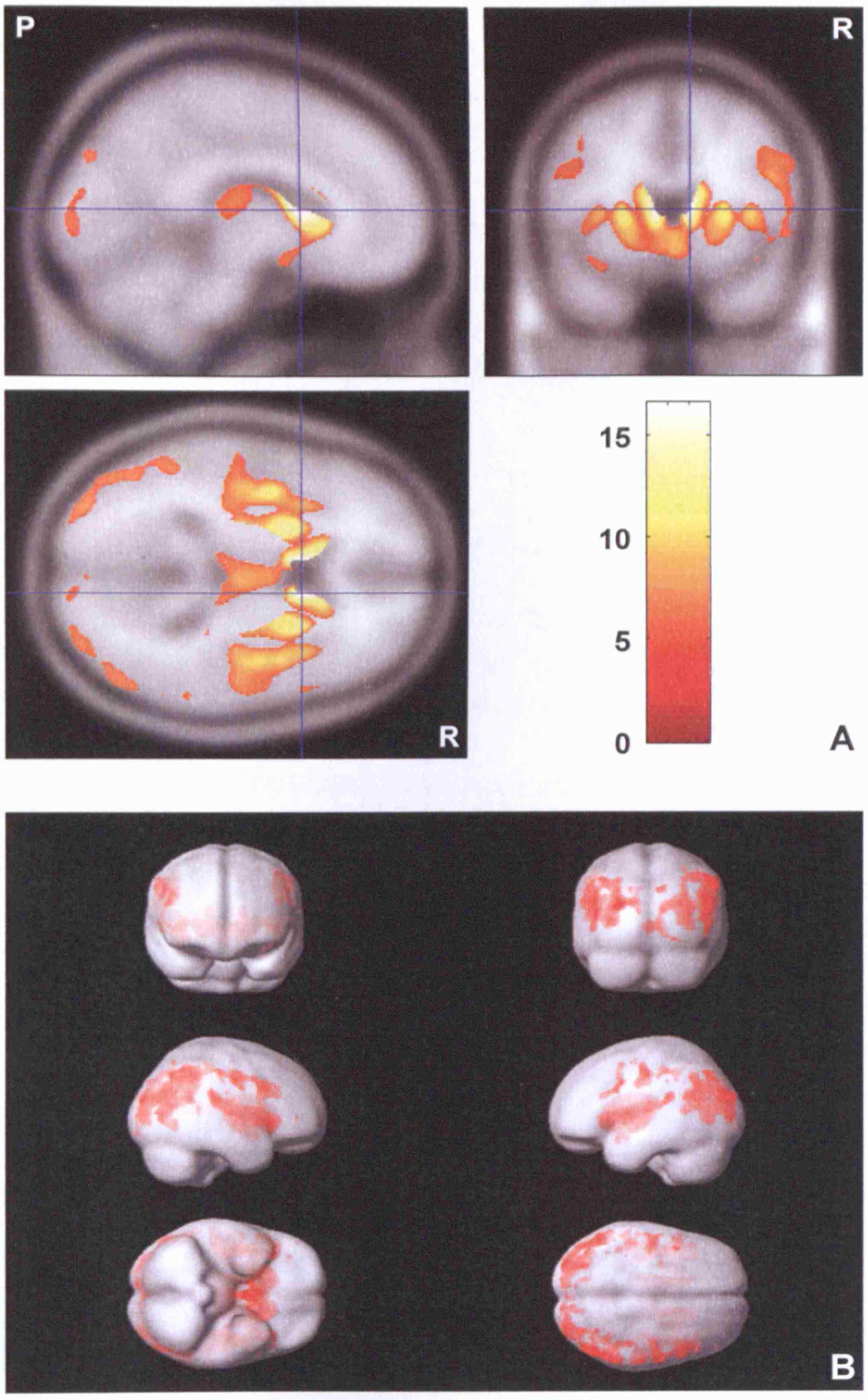


Figure 7-1 A) Regions of reduced grey matter in early HD relative to controls; the colour bar shows the t score (coordinates 9, 10, 9mm); B) the same contrast rendered on the MNI-152 average brain to show cortical atrophy; p<0.05 FWE

Table 7-1 Location of maxima for regions in which early HD subjects had significantly reduced GM volume relative to controls, FWE corrected, p<0.05

Region	L / R	MNI coordinates			p value
		x	y	z	
Caudate	L	-7	8	8	<0.0001
Caudate	R	9	10	9	<0.0001
Putamen	L	-26	1	11	<0.0001
Insula	L	-42	-5	8	<0.0001
Insula	R	42	-11	10	<0.0001
Thalamus	L	-10	-21	16	0.001
Thalamus	R	12	-25	13	0.001
Inferior frontal gyrus	L	-43	7	38	<0.0001
Medial frontal gyrus	L	-44	6	52	0.001
Medial frontal gyrus	R	49	44	3	0.01
Precentral gyrus	L	-53	4	35	0.001
Precentral gyrus	R	50	-13	56	0.036
Antero-medial temporal lobe	L	-35	9	-19	0.001
Inferior temporal gyrus	L	-51	-65	-18	<0.0001
Superior temporal gyrus	L	-64	-33	16	0.027
Superior temporal gyrus	R	60	-43	11	0.013
Angular gyrus	R	53	-64	39	<0.0001
Inferior parietal lobule	L	-43	-44	43	0.002
Postcentral gyrus	L	-45	-17	36	<0.0001
Postcentral gyrus	R	61	-18	27	0.004

Region	L / R	MNI coordinates			p value
		x	y	z	
Superior parietal lobule	L	-40	-56	54	0.025
Supramarginal gyrus	L	-60	-51	35	0.042
Medial occipital gyrus	L	-33	-92	6	<0.0001
Superior occipital gyrus	L	-25	-87	22	<0.0001
Superior occipital gyrus	R	36	-83	20	<0.0001
Cuneus	R	10	-86	36	<0.0001
Cingulate gyrus	R	3	27	17	0.029

SPM smoothness 9.4, 10.2, 9.5mm, 1773.6 resels (see section 2.5.5); FWE = family-wise error

7.3.2 *Early HD vs. Premanifest*

The early HD group also had significantly less grey matter than the premanifest group. The pattern of atrophy was similar to above but reduced in extent and mainly restricted to small regions of the striatum and insula; there was much less extreme atrophy in the cortical areas mentioned above (Table 7-2, Figure 7-2). There were no significant differences between the controls and premanifest group. There were no regions of significant increase in premanifest subjects compared to controls.

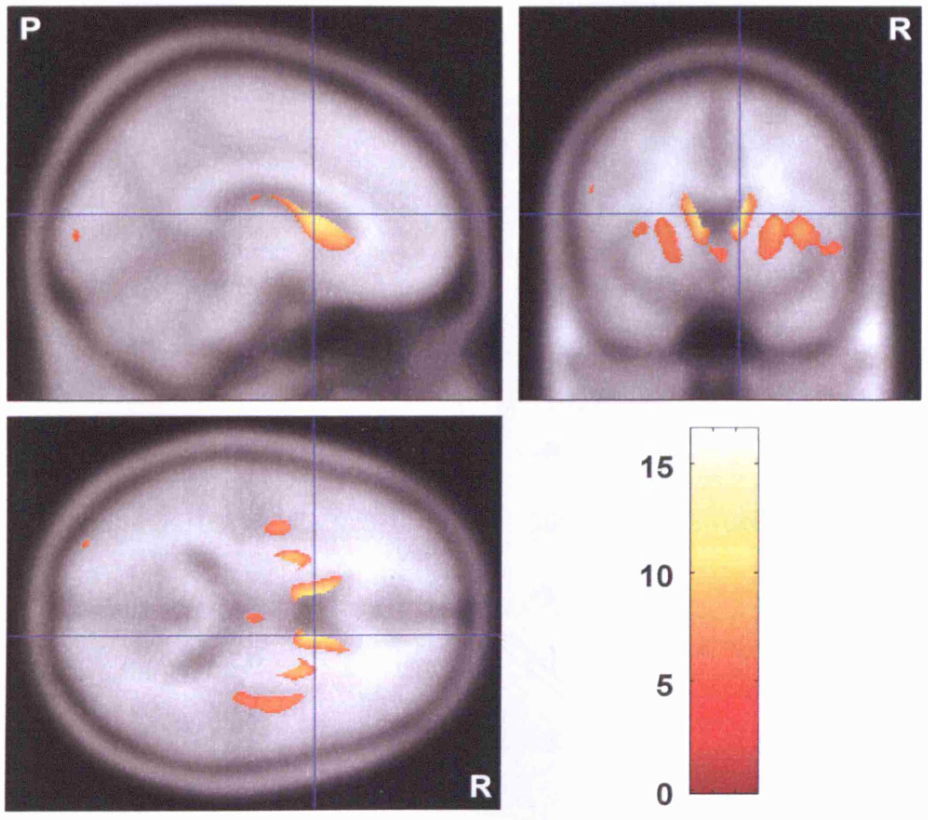


Figure 7-2 Regions of reduced grey matter in early HD relative to premanifest subjects (coordinates 10, 9, 13mm); p<0.05 FWE

Table 7-2 Location of maxima for regions in which early HD subjects had significantly reduced GM volume relative to premanifest, FWE corrected, p<0.05

Region	L / R	MNI coordinates			p value
		x	y	z	
Caudate	L	-9	9	13	<0.0001
Caudate	R	10	9	13	<0.0001
Putamen	L	-24	-1	11	<0.0001
Putamen	R	27	6	8	<0.0001
Insula	L	-42	-5	7	<0.0001
Insula	R	37	11	5	<0.0001
Medial frontal gyrus	L	-26	-8	57	0.023
Inferior frontal gyrus	L	-58	10	25	0.024
Precentral gyrus	R	61	-11	39	0.005
Superior temporal gyrus	L	-45	14	-7	0.02
Medial temporal gyrus	L	-46	-77	23	0.002
Medial occipital gyrus	L	-36	-89	6	0.001
Medial occipital gyrus	R	35	-83	19	0.004
Inferior occipital gyrus	R	15	-97	-11	0.032
Cuneus	L	-13	-98	18	0.033
Cuneus	R	21	-91	23	0.001

SPM smoothness 9.4, 10.2, 9.5mm, 1773.6 resels; FWE = family-wise error

7.3.3 *HD vs. Controls, adjusting for total grey matter volume*

After controlling for total grey matter loss, the early HD group still had significantly less grey matter than controls; the caudate, putamen and insula bilaterally show focal atrophy disproportionately greater than the overall volume loss (Table 7-3, Figure 7-3). There were no areas in which there was significant relative sparing of grey matter relative to controls.

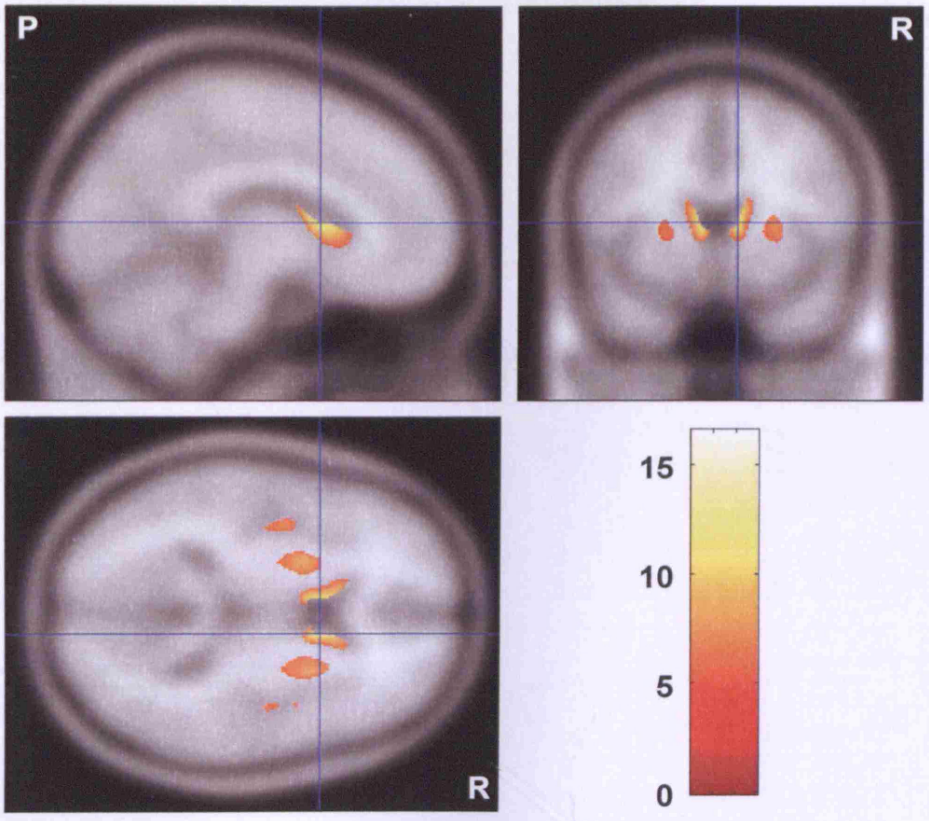


Figure 7-3 Regions of reduced grey matter in HD relative to controls after controlling for global disease effects (coordinates 9, 10, 9mm); p<0.05 FWE

Table 7-3 Location of maxima for regions in which early HD subjects had significantly reduced GM volume relative to controls after controlling for global atrophy, FWE corrected, p<0.05

Region	L / R	MNI coordinates			p value
		x	y	z	
Caudate	R	9	10	9	<0.0001
Caudate	L	-7	8	8	<0.0001
Putamen	R	26	8	8	<0.0001
Putamen	L	-26	1	11	<0.0001
Insula	R	39	-19	15	0.007
Insula	L	-42	-5	8	<0.0001
Superior occipital gyrus	L	-24	-87	22	0.001

SPM smoothness 9.2, 9.9, 9.2mm, 1927.6 resels; FWE = family-wise error

7.3.4 *Effect of CAG repeat length*

After controlling for age, gender and head size (TIV), higher CAG repeat length was significantly associated with reduced volume in the body of the caudate nucleus bilaterally, left putamen, right insula, right parahippocampal gyrus, right anterior cingulate, and right occipital lobe (Figure 7-4). CAG-related atrophy was more extensive on the right than on the left. There were no regions in which higher CAG repeat length was significantly associated with less atrophy.

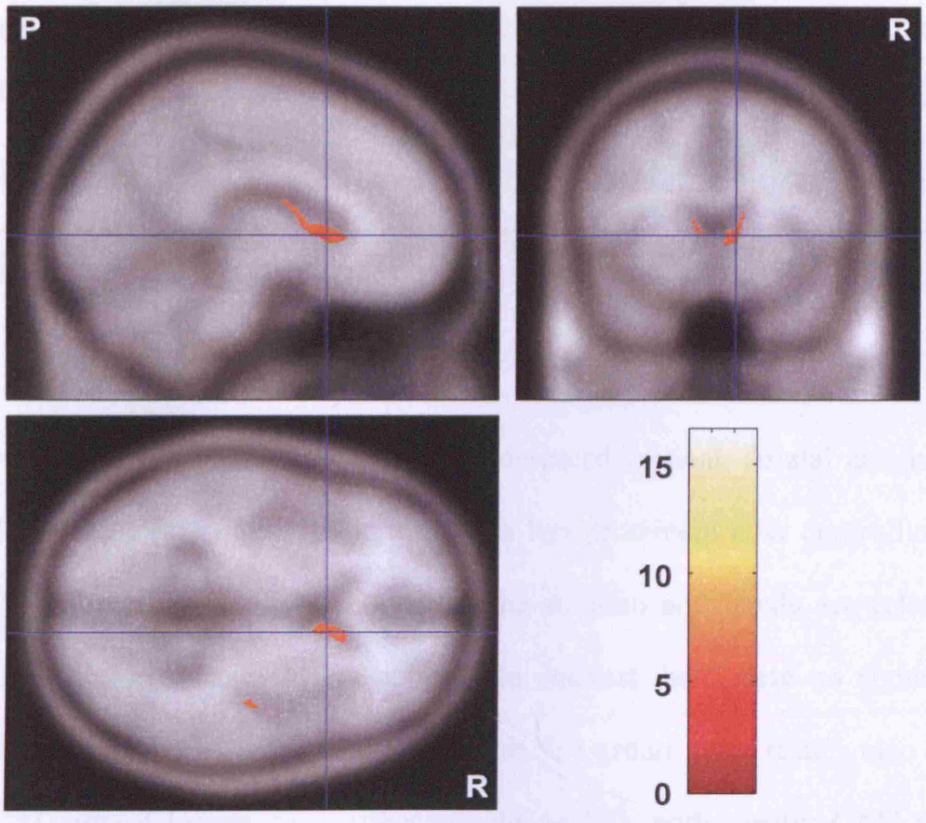


Figure 7-4 Regions in which grey matter decreases with increased CAG repeat length (coordinates 10, 17, 8mm); p<0.05 FWE

Table 7-4 Location of maxima for regions in which grey matter decreased with increased CAG repeat length, FWE corrected, p<0.05

Region	L / R	MNI coordinates			p value
		x	y	z	
Caudate	L	-9	5	13	<0.0001
Caudate	R	10	17	8	<0.0001
Putamen	L	-24	-4	11	0.009
Insula	R	44	0	7	0.045
Parahippocampal gyrus	R	14	2	-18	<0.0001
Anterior cingulate	R	1	4	-13	0.001
Inferior occipital gyrus	R	25	-71	-16	0.048

SPM smoothness 9.5, 10.4, 9.6mm, 1702.8 resels; FWE = family-wise error

7.4 DISCUSSION

This work has confirmed that there is widespread cortical, striatal and insular atrophy in early HD. Cortical atrophy was less prominent after controlling for global grey matter loss, suggesting that the striatum and insula are relatively more affected in this early HD cohort. In contrast there were no significant differences between controls and the premanifest group. The results also show that CAG repeat length does affect atrophy in HD, with greater CAG repeat length associated with greater loss of caudate, insula and cortical volume.

7.4.1 Volume changes in HD

Striatal atrophy is well documented in HD, and the current study confirmed this. In addition atrophy was also found in the insula bilaterally and the thalamus, in contrast to some previous VBM studies (Kassubek *et al.* 2004c; Kassubek *et al.*

2005). The presence and extent of cortical atrophy in HD also varies between studies with some reporting very little extra-striatal atrophy outside the insula, whilst others report motor, somatosensory or fronto-parietal cortical changes (Kassubek *et al.* 2004c; Peinemann *et al.* 2005; Douaud *et al.* 2006). Of particular relevance to this study is the finding by Mühlau *et al.* (2007b) of more extensive atrophy involving parietal, occipital, and frontal regions, a pattern of change which was also found here. Techniques such as cortical thickness measurement support this, showing regions of cortical thinning including occipital, temporal and (in more advanced stages) frontal lobes (Rosas *et al.* 2002). This fits well with post-mortem data, where there is a >20% reduction in cortical grey matter, compared with controls (de la Monte *et al.* 1988; Halliday *et al.* 1998).

This evidence that cortical atrophy is quite widespread – even early in the disease – has not been consistently reported in the HD imaging literature. It is plausible that cortical effects, if less severe than striatal atrophy, are failing to reach significance owing to insufficient power in studies with smaller sample sizes. When I examined the contrast maps[1] for the differences between HD and controls these differences were largest in the striatum, possibly because the striatum starts to atrophy prior to cortical regions, or because it atrophies at a faster rate.

---

[1] The contrast map is the numerator of the t-statistic (it shows the difference between the parameter estimates before being divided by the pooled standard deviation) and is hence a measure of the magnitude of the effect.

There is some evidence of abnormal cortical volume in premanifest subjects, ranging from decreased cortical thickness and grey matter volume (Thieben *et al.* 2002; Rosas *et al.* 2005; Kipps *et al.* 2007), to increased total grey matter volume (Paulsen *et al.* 2006b). However, in keeping with the findings of Wolf *et al.* (2007), in my sample of 21 premanifest subjects I found no significant regions of reduced grey matter compared with controls, although the group did include a number (57%) who were >15 years prior to predicted onset. Whole-brain volumes in this group were slightly smaller than controls (see section 4.3.3), so although there may be some volume loss it is likely to be small this many years prior to onset.

Finally there is the interesting issue of asymmetry of these cortical changes. A recent VBM study reported leftward bias in atrophy in HD (Mühlau *et al.* 2007a) but in this cohort such asymmetry was not obvious from visual inspection of the SPM (in contrast to Mühlau *et al.*) although there were differences in some structures between the left and the right hemisphere.

7.4.2 CAG repeat length and volume changes in HD

Higher CAG repeat length was associated with reduced volume in the body of the caudate bilaterally, and the right insula, as well as small regions in the anterior cingulate, temporal, and inferior occipital lobes (after adjusting for age). Most other ROI studies have found a relationship between CAG repeat length and striatal volume in manifest subjects (Rosas *et al.* 2001; Ruocco *et al.* 2006). In my study the effect was smaller (and less widespread) in the putamen than the caudate, which may reflect a quantitative difference in that the effect of repeat length on caudate volume is simply larger than that on the putamen volume, as

the medium spiny neurons in the caudate are known to degenerate preferentially in HD (Wanker and Droge 2002). It is also not clear whether atrophy in extra-striatal regions is secondary to striatal atrophy or represents an independent parallel degenerative process, but given that CAG repeat length varies across brain regions (Shelbourne *et al.* 2007) it is possible that independent parallel processes of degeneration occur at different CNS sites, and that these may also account for some of the early non-motor features of the disease.

In HD brain volume will be affected by the rate and duration of the pathological process, as well as age (because of the normal loss from healthy ageing). Assuming that the pathological process in HD starts some time prior to symptom manifestation rather than progressing linearly from birth (Rosenblatt *et al.* 1998) then for people of the same age, those with a higher number of CAG repeats are likely to have had earlier symptom onset and hence a longer pathological process. There is also some evidence that higher CAG repeat length causes faster rates of progression (Aylward *et al.* 1997; Rosenblatt *et al.* 2006). However, it is not possible to determine the length of time for which the pathological process has been occurring; duration of motor symptoms is a very approximate guide, often determined retrospectively and in any case not applicable to premanifest subjects. Therefore it is not possible to conclude whether the effect of CAG repeat length seen here is due to differences in the duration or rate of the pathological process. However it is clear that both striatal and extra-striatal atrophy are predicted by CAG repeat length. Future work with prospective longitudinal whole-brain studies need to be undertaken to explore the regional change in atrophy rates and their relationship to CAG repeat length.

7.4.3 Technical implications

There are several explanations for the differences between these findings and those of previous studies. One reason for the lack of consensus may be that HD is clinically a highly variable disease, and so conflicting findings reflect real differences in the populations being studied. However the mean ages, disease durations and CAG repeat lengths (where specified) of the early HD cohorts in other studies are not very different to those in the current study (Kassubek *et al.* 2004c; Peinemann *et al.* 2005; Mühlau *et al.* 2007b). One explanation that must therefore be considered is variations in methodology; the techniques used in this study have been successfully employed in other studies in patients with neurodegenerative disorders, but, as discussed in the previous chapter, differences in processing parameters and statistical models may have relatively large effects on findings. Many of the previous studies using VBM in HD have used some form of "optimised" VBM which is likely to have improved normalisation relative to standard VBM. However, some did not modulate data, smoothing kernels were variable, few studies considered the effects of head size or gender, and the resulting SPMs were thresholded at widely varying levels using different types of correction, all of which are likely to have contributed to the disparity between findings.

The work in the previous chapter aimed to maximise the accuracy of the SPM2 normalisation and segmentation algorithms for this dataset, and models included age, head size and gender as covariates. Results were only considered statistically significant after controlling the FWE rate, so false positives are unlikely, although there is a caveat that some true positives are undoubtedly missed. Overall this work aimed to minimise sources of error from VBM

preprocessing, through to model-fitting and statistical inference, meaning that the effects seen here are likely to be relatively robust.

7.5 CONCLUSION

Voxel-based morphometry is becoming increasingly used as a tool to investigate patterns of atrophy in HD, but it is hard to generalise results because of the many different options that can be set by the operator. Despite this there are a number of common findings between this study and other studies using both VBM and different techniques, which allows one to draw some conclusions regarding the pattern of atrophy in HD. In my large cohort I used a whole-brain analysis (rather than predefined regions of interest) with a strict correction level and found evidence for atrophy of the striatum and insula as well as a number of cortical areas. I have also shown that CAG repeat length is related to volume in the caudate nucleus, putamen and some regions of the cortex, in a sample of 61 mutation-positive HD subjects (21 premanifest, 40 early). This has important implications for research and in particular studies assessing disease-modifying therapies, as CAG repeat length will affect the pattern and extent of atrophy. The finding of cortical atrophy even in the early stages reinforces the idea that striatal damage is not the only underlying cause of many of the deficits seen in HD and probably occurs in parallel and independently of the basal ganglia pathology.

8 THE NEURAL BASIS OF THE EMOTION RECOGNITION DEFICIT IN HUNTINGTON'S DISEASE

8.1 INTRODUCTION

As discussed in chapter 1, deficits in emotion recognition have been reported in premanifest and early HD, and been shown in the visual, vocal, olfactory and gustatory domains (Sprengelmeyer *et al.* 1996; Gray *et al.* 1997; Mitchell *et al.* 2005; Hayes *et al.* 2007). However although a number of studies found that disgust recognition was disproportionately impaired, others have failed to replicate this (Milders *et al.* 2003; Johnson *et al.* 2007). As with other discrepancies in the literature, this may be attributable to the heterogeneity of the disease, but these authors also point out analytic differences that might have influenced results.

The presence of focal striatal atrophy in HD (Vonsattel *et al.* 1985) has often led to suggestions that striatal damage might underlie the impairment in disgust recognition, and recently functional and structural imaging studies have found evidence for striatal involvement in recognition of disgust in both visual and auditory modalities (Sprengelmeyer *et al.* 1998). However, the specificity of this association has not been established. In functional imaging studies the anterior insula is also implicated in recognition of disgust (Phillips *et al.* 1997; Phillips *et al.* 1998; Sprengelmeyer *et al.* 1998; Murphy *et al.* 2003; Hennenlotter *et al.* 2004; Kipps *et al.* 2007), while striatal activation is also associated with fear recognition (Phillips *et al.* 1998). These findings suggest that the striatum and insula may jointly participate in processing disgust and other negative emotions. Functional imaging evidence is reinforced both by depth electrode studies of

disgust processing (Krolak-Salmon *et al.* 2003), and by human lesion studies (Calder *et al.* 1996; Sprengelmeyer *et al.* 1997; Calder *et al.* 2000; Calder *et al.* 2004).

The work in this chapter was therefore carried out in order to clarify the emotion recognition deficit in early HD and its brain basis by studying facial emotion recognition in a large, well-defined disease cohort. Voxel-based morphometry (VBM) (Ashburner and Friston 2000) was used to examine associations between emotion recognition performance and grey matter atrophy, both across the whole brain and in specific regions of interest suggested by previous research. It was predicted that a distributed network of structures would be involved including the striatum, insula, amygdala and orbitofrontal cortex, as these regions have been widely implicated in emotion recognition both in HD and other disease states and in healthy subjects (Sprengelmeyer *et al.* 1998; Blair *et al.* 1999; Kipps *et al.* 2007).

8.2 METHODS

8.2.1 Subjects

Subjects were the same 81 as those reported in chapter 4, see section 4.2.1 for details.

8.2.2 Assessments

Subjects underwent the same assessments described in chapter 4, sections 4.2.2.1, 4.2.2.2, and 4.2.2.3. That is, they underwent clinical and cognitive assessments and had 1.5T MRI scans. Of the large cognitive battery reported in chapter 4, scores for Recognition Memory for Faces (RMF), the silhouette

subtest of the VOSP, the Benton Facial Recognition Test and the Beck Depression Inventory (BDI) are reported again here as the skills they tap are relevant to the main skill under investigation, facial emotion recognition.

Subjects were given an emotion recognition test as described in section 4.2.2.2. Twenty-four faces from the Ekman and Friesen battery were used (Ekman and Friesen 1976; see Gray *et al.* 1997). This set was chosen so that the resulting dataset would be comparable to that of Gray *et al.* who were able to demonstrate a selective impairment in disgust recognition using the same set on a cohort of premanifest HD subjects. It was also shorter than the set of 60 faces or morphs sometimes used, which made it more suitable for the assessment of manifest subjects. Faces were displayed on individual cards and presented to subjects in a random order. The words "Happiness", "Sadness", "Surprise", "Disgust", "Anger", "Fear" were printed beneath each face in a pseudo-random order, and subjects were instructed to look at the face, read the choices, and decide which of the words best described the emotion shown by the face. Six practice faces were administered beforehand. There was no time limit.

8.2.3 Statistical Analysis

Linear regression models were used to compare mean levels on the VOSP, RMF, Beck and Benton between groups. The effects of variables which could independently affect cognitive performance were controlled for by including them as covariates where necessary. Age and estimated premorbid IQ were included as covariates for variables other than the Beck (as mood is not known to vary with age and IQ) and the Benton, which is already adjusted for age and education.

Linear regression models were also used to compare mean levels on each emotion score from the Ekman data. In addition to age and estimated premorbid IQ, scores on the Benton test were included as a covariate in this model to adjust for any differences in the perceptual analysis of faces. As there were only five possible scores per emotion, and consequently Normality assumptions were seriously violated, a 95% bias-corrected bootstrap confidence interval with 1000 replicates was used, here and in later analyses.

In order to allow for the fact that some emotions are harder to recognise than others, scores were standardised by subtracting the control mean from each score and dividing this difference by the control standard deviation (i.e. converted to z-scores). Because of the discrete nature of the data the resultant z-scores were not Normally distributed. For this reason Wilcoxon signed-rank tests were used to look for differences between the z-scores for disgust and the other two worst recognised emotions (anger and fear). Z-scores were not calculated for happy as the control group scored at ceiling and thus had a standard deviation of zero. A 95% bias-corrected bootstrap confidence interval with 1000 replicates was also used to compare control data to Ekman's published mean levels.

8.2.4 VBM analysis

Images were preprocessed following the steps described in section 2.5.5.

Scores on the Ekman facial emotion recognition test were entered into linear regression models to look at group by score interaction, with the exception of happiness on which most subjects were at ceiling. For each emotion, voxel intensity, V, was modelled as a function of group and score and their interaction,

controlling for age, gender, TIV, IQ and Benton facial recognition score by including them in the model as covariates (equation 8.1).

$$V = \beta_1 \text{ Control} + \beta_2 \text{ HD gene} + \beta_3 \text{ Control} \times \text{score} + \beta_4 \text{ HD} \times \text{score} + \beta_5 \text{ age} + \beta_6 \text{ gender} + \beta_7 \text{ TIV} + \beta_8 \text{ IQ} + \beta_9 \text{ Benton} + \mu + \varepsilon \quad (8.1)$$

This model allows investigation of the association between GM volume and score in HD subjects and/or controls ($\beta_4 > 0$ or $\beta_3 > 0$), and also to test whether the magnitude (slope) of the association differs between groups ($\beta_4 > \beta_3$) thus suggesting a disease-related finding. TIV was measured according to the protocol described in Appendix 2.

For these analyses all HD gene carriers were included as a single group (N=61) since the onset of motor signs is not necessarily associated with the onset (and therefore severity) of cognitive signs, and this provided the maximum power to detect relationships between score and atrophy. A number of studies have now shown that both atrophy and reduced emotion recognition performance can be detected prior to motor onset (Aylward *et al.* 1994; Johnson *et al.* 2007).

Each emotion was examined at the whole-brain level, and then in pre-specified ROIs based on existing literature. Anatomical small volumes for the orbitofrontal cortex, right insula and striatum were created in MRIcro (http://www.sph.sc.edu/comd/rorden/mricro.html) by manually outlining ROIs on a control brain in MNI space, with reference to an anatomical map. The left insula ROI was a mirror image of the right insula ROI. The amygdala was segmented on a healthy control by an image analyst at the DRC using semi-automated software (Freeborough *et al.* 1997) and this was put into standard

space in SPM2 and converted into a binary mask. All ROIs were smoothed at 2mm FWHM and thresholded at 0.25. ROI analysis was performed by selecting "Small Volume Correction" in SPM2 and then applying the particular region of interest. ROIs are shown in Figure 8-1.

The resulting SPMs were masked as described in section 2.5.5.2 and thresholded at $q<0.05$ using FDR correction (Genovese *et al.* 2002) (see also section 6.2).

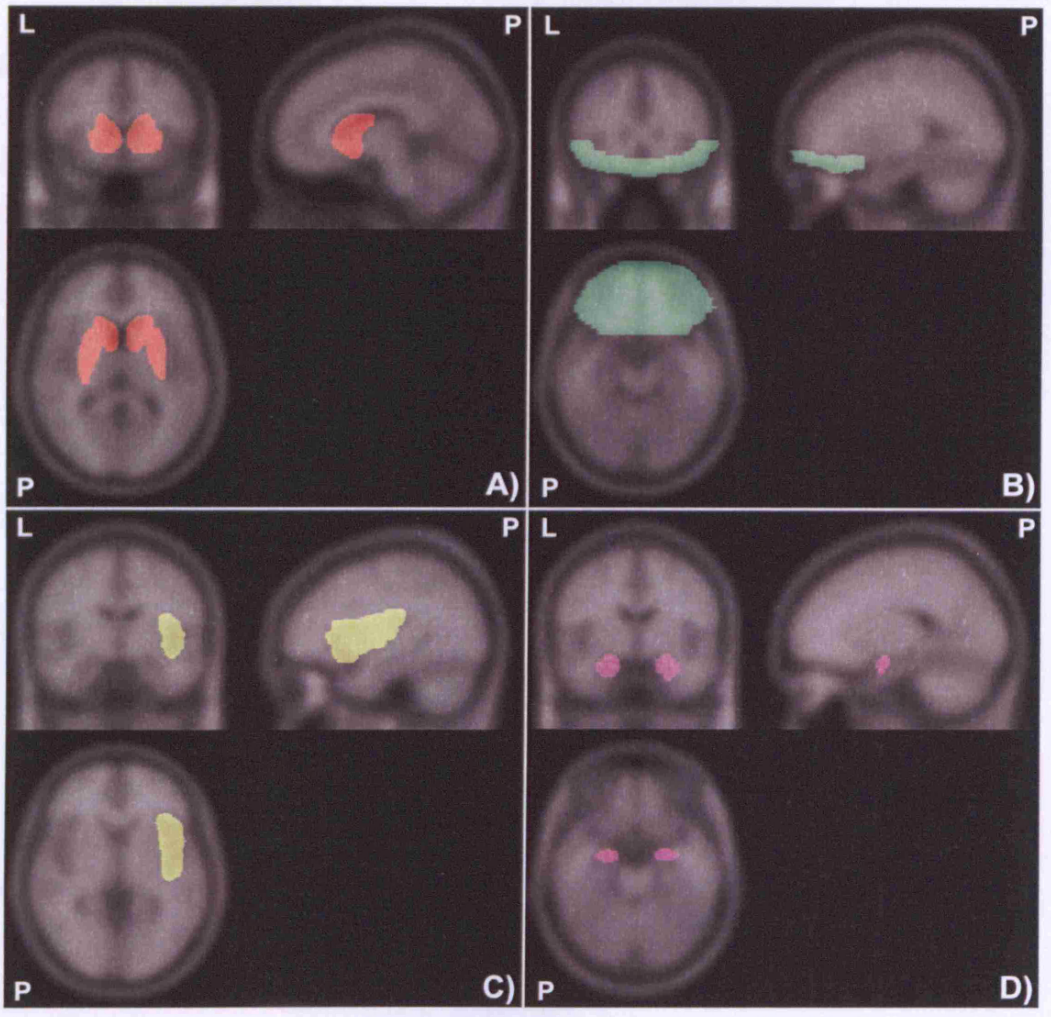


Figure 8-1 Regions of interest (ROIs) used in the small volume analysis. A) striatum, B) orbitofrontal cortex, C) right insula (left also tested), D) amygdala

The association between volume and score for each emotion was investigated within these five ROIs.

8.3 RESULTS

As subjects were the same as those reported in chapter 4, demographic data can be found in Table 4-1, page 135.

8.3.1 Cognition

There was no evidence of differences between the groups on the Beck and VOSP silhouette test but the HD group was worse at the Benton than both controls ($p<0.0001$) and premanifest ($p<0.0001$) and at the RMF (HD vs. controls, $p<0.0001$, HD vs. premanifest, $p=0.001$), Table 8-1.

Table 8-1 Mean (SD) cognitive scores, and differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ

	Control	Premanifest	Early HD	Difference (Premanifest - Control)		Difference (Early HD - Control)	
	(N=20)	(N=21)	(N=40)	Crude	Adjusted	Crude	Adjusted
BDI	5.6 (3.9)	6.8 (6.3)	9.3 (8.7)	1.3 (-3.2, 5.7)		3.7 (-0.2, 7.7)	
				p=0.57		p=0.06	
VOSP (/15)	9.5 (2.3)	8.6 (3.2)	8.6 (2.3)	-0.8 (-2.4, 0.8)	-1.1 (-2.7, 0.6)	-0.8 (-2.2, 0.6)	-0.6 (-2.0, 0.8)
				p=0.31	p=0.20	p=0.25	p=0.40
Benton (/54)	48.0 (3.8)	47.7 (3.7)	43.5 (4.7)	-0.3 (-2.9, 2.4)		-4.5 (-6.8, -2.2)	
				p=0.83		p<0.0001	
RMF (/50)	42.7 (4.0)	41.3 (5.3)	34.8 (7.0)	-1.4 (-5.1, 2.3)	-1.5 (-5.2, 2.3)	-7.9 (-11.2, -4.7)	-7.5 (-10.7, -4.3)
				p=0.47	p=0.44	p<0.0001	p<0.0001

Negative differences indicate a lower score in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models; the BDI was not adjusted for age and IQ, and the Benton scores are already age- and education-adjusted.

After adjusting for age, IQ and Benton facial recognition score the premanifest group was significantly worse than controls at recognising happiness ($p<0.05$), although there was no evidence of a difference between the two patient groups for this emotion. The early HD group was significantly worse than controls at recognising facial expressions of surprise, disgust, anger and fear (all $p<0.05$) and worse than premanifest subjects at recognising disgust and anger (all $p<0.05$, Table 8-2, Figure 8-2).

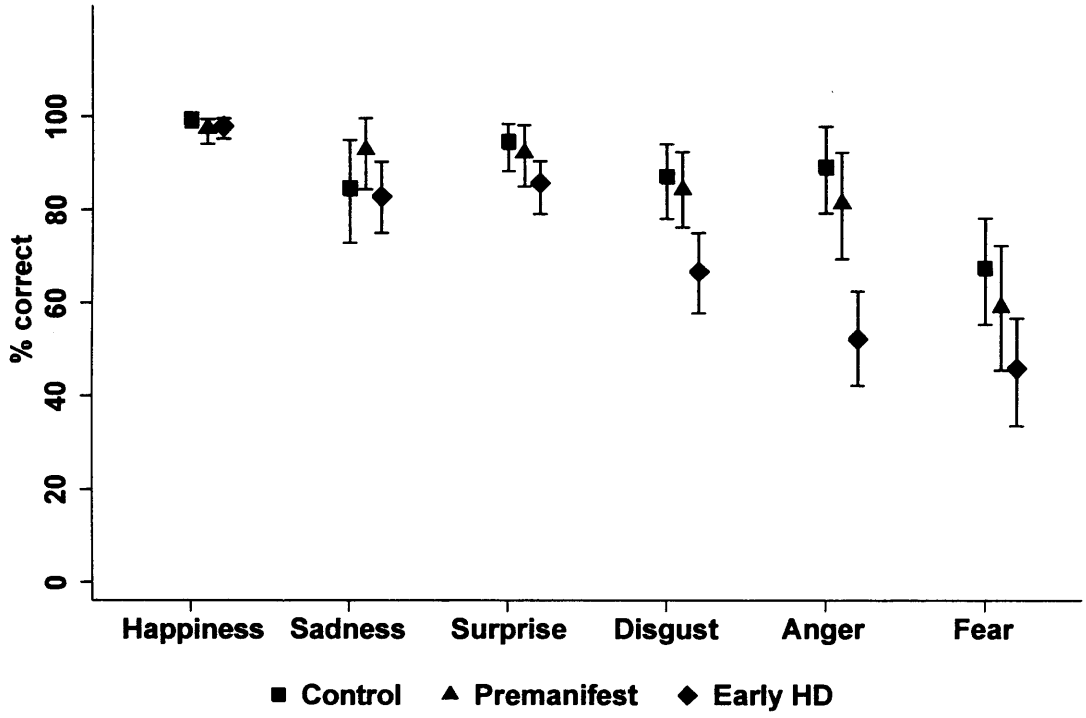


Figure 8-2 Mean emotion score (% correct) for each group with 95% bias-corrected bootstrap confidence intervals calculated from 1000 replications

Table 8-2 Mean (SD) % correct for Ekman Pictures of Facial Affect with differences (95% confidence intervals) with and without adjustment for age, estimated premorbid IQ and Benton score

	Control	Premanifest	Early HD	Difference (Premanifest - Control)		Difference (Early HD - Control)	
	(N=20)	(N=21)	(N=40)	Crude	Adjusted (95% CI)	Crude	Adjusted (95% CI)
Happiness	100 (0)	98.8 (5.5)	96.9 (8.4)	-1.2 (-4.8, 0.0)	**-1.8** (-4.6, -0.1)	**-3.1** (-6.3, -0.7)	-1.4 (-3.9, 0.7)
Sadness	86.3 (22.2)	92.9 (14.0)	81.9 (21.2)	6.6 (-3.6, 18.2)	8.4 (-4.4, 21.5)	-4.4 (-15.0, 8.3)	-1.7 (-15.9, 11.7)
Surprise	95.0 (10.3)	92.9 (14.0)	85.0 (15.8)	-2.1 (-9.7, 4.8)	-2.3 (-11.1, 5.1)	**-10.0** (-17.1, -3.6)	**-8.7** (-16.6, -1.6)
Disgust	90.0 (12.6)	82.1 (21.1)	66.3 (30.8)	-7.9 (-19.0, 1.1)	-2.7 (-12.6, 7.6)	**-23.8** (-35.0, -13.5)	**-20.3** (-31.0, -8.7)
Anger	91.3 (14.7)	84.5 (24.3)	49.4 (32.3)	-6.7 (-19.6, 5.0)	-7.7 (-21.8, 6.3)	**-41.9** (-53.0, -29.3)	**-36.8** (-52.1, -22.4)
Fear	71.3 (23.3)	64.3 (31.2)	41.3 (32.8)	-7.0 (-26.3, 8.0)	-8.3 (-24.2, 8.6)	**-30.0** (-43.7, -15.0)	**-21.4** (-37.4, -4.3)

Negative differences indicate a lower score in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age, IQ and Benton score; adjusted differences are the differences adjusting for age, IQ and Benton by including these as covariates in the linear regression models; confidence intervals are 95% bias-corrected bootstrap confidence intervals calculated from 1000 replications; differences shown in bold are significant at $p<0.05$.

When emotion recognition scores were expressed as z-scores in order to control for emotion difficulty, subjects with HD were worse at recognising anger compared with disgust (p=0.023) and anger compared with fear (p=0.0006). The difference between disgust and fear was not statistically significant (Figure 8-3).

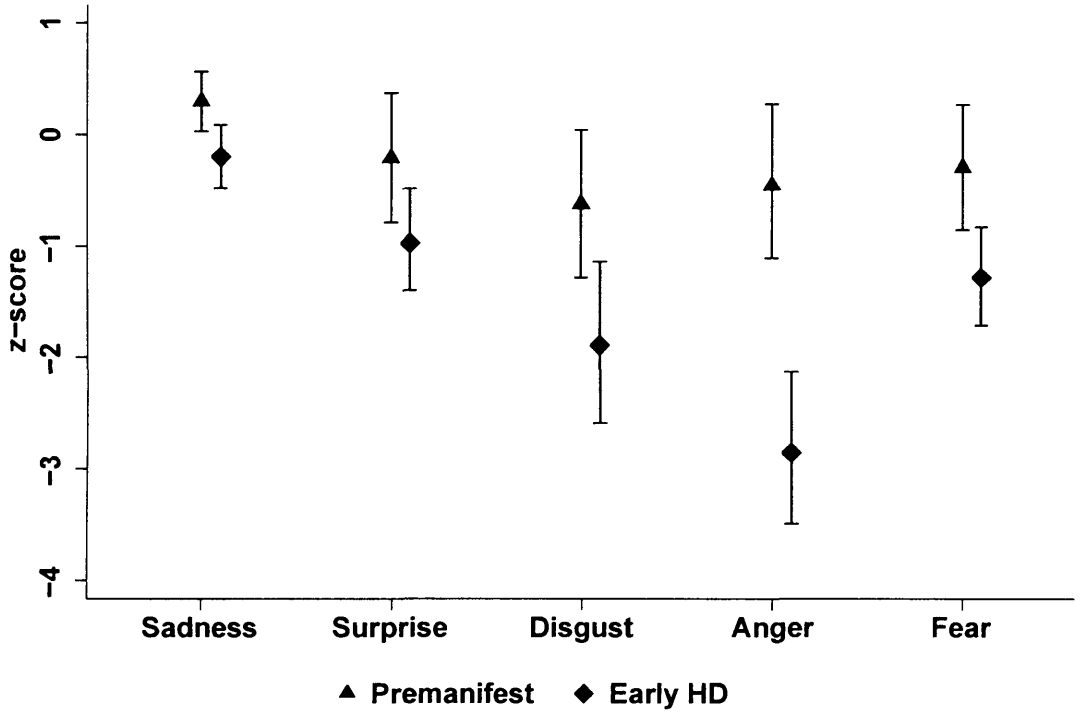


Figure 8-3 Premanifest and early HD z-scores with 95% bias-corrected bootstrap confidence intervals calculated from 1000 replications

The control group was statistically significantly worse than the Ekman published norms at recognising sadness, anger and fear (all $p<0.05$, Table 8-3).

Table 8-3 Control mean % correct with 95% bootstrap confidence intervals and Ekman norms

	Control (N=20)	Ekman
Happiness	100 (100)	100.0
Sadness	**86.3** (76.3, 95.0)	95.8
Surprise	95.0 (91.3, 100)	95.8
Disgust	90.0 (85.0, 96.3)	96.0
Anger	**91.3** (85.0, 97.5)	98.3
Fear	**71.3** (62.5, 81.3)	93.0

Means shown in bold are significantly different to the Ekman norms, $p<0.05$.

8.3.2 Neuroanatomical correlates

8.3.2.1 Whole-brain analyses

When associations between emotion recognition score and grey matter volume were examined across the whole brain none were significant after correction for multiple comparisons. However the effect maps are shown in Figure 8-4 in order to demonstrate how the associations varied both across the brain and between emotions. Results for the ROI analyses are reported below, with the exception of sadness, for which there was no evidence of significant associations (see also Table 8-4).

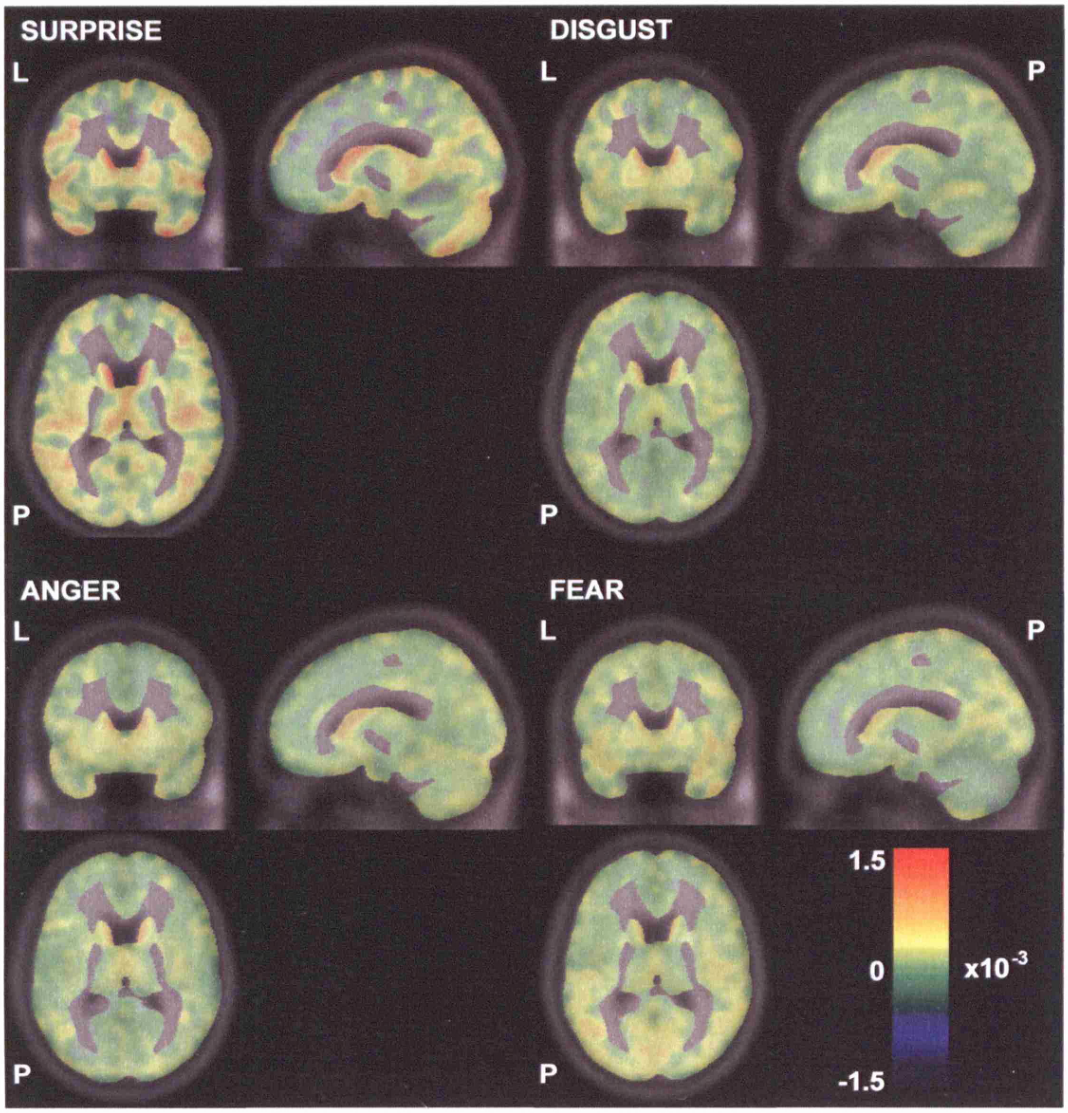


Figure 8-4 Effect maps for the associations between grey matter volume and surprise, disgust, anger and fear

Colour bar represents the size of the effect from positive (red) through zero (green) to negative (blue) (coordinates 14, 14, 12mm). This maps the strength of the association between volume and emotion score at each voxel (rather than the significance of that effect). Red indicates a strong positive association (larger volume = larger score), green indicates no association and blue indicates a negative association.

8.3.2.2 Surprise

In the HD gene carriers (N=61) impaired recognition of surprise was associated with atrophy in the caudate body bilaterally (Figure 8-5). There were no significant associations in the other ROIs or in the control group.

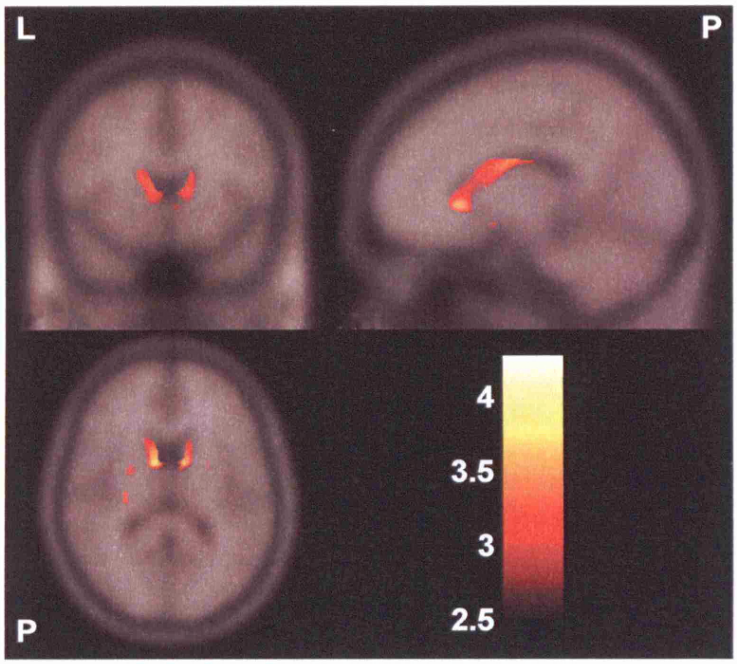


Figure 8-5 Regions in which surprise performance was associated with grey matter atrophy in the striatal ROI (coordinates 14, 14, 12mm); q<0.05 FDR, smoothness 9.5, 10.3, 9.6mm, 54.7 resels; colour bar shows t score

8.3.2.3 Disgust

In the HD gene carriers deficits in disgust recognition were associated with grey matter atrophy throughout the striatal ROI, with peaks in the left caudate head and caudate body bilaterally (Figure 8-6). There were no significant associations in the other ROIs or in the control group.

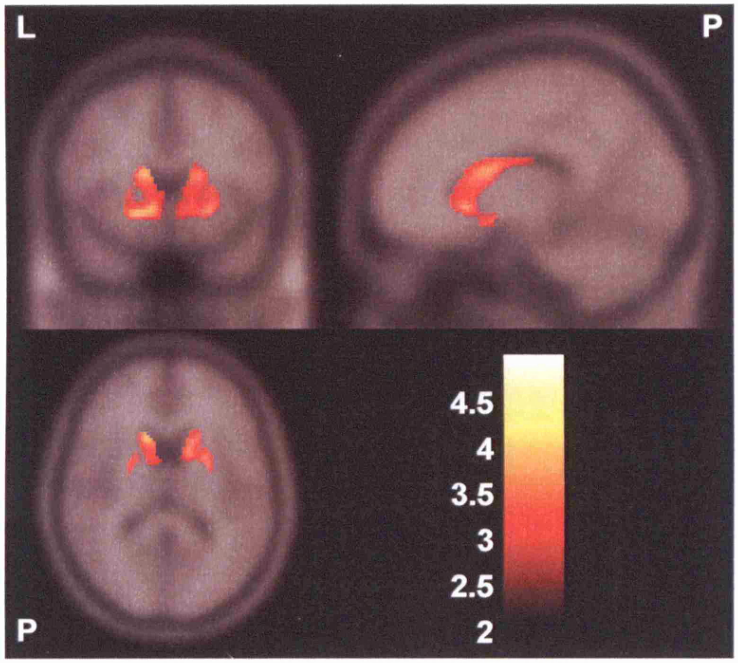


Figure 8-6 Regions in which disgust performance was associated with grey matter atrophy in the striatal ROI (coordinates 14, 14, 12mm); q<0.05 FDR, smoothness 9.5, 10.3, 9.6mm, 54.4 resels

8.3.2.4 Anger

Deficits in anger recognition in the HD gene carriers were also associated with grey matter atrophy throughout the striatal ROI, with peaks in the head and body of the caudate bilaterally, and the right putamen (Figure 8-7). There were no significant associations in the other ROIs or in the control group.

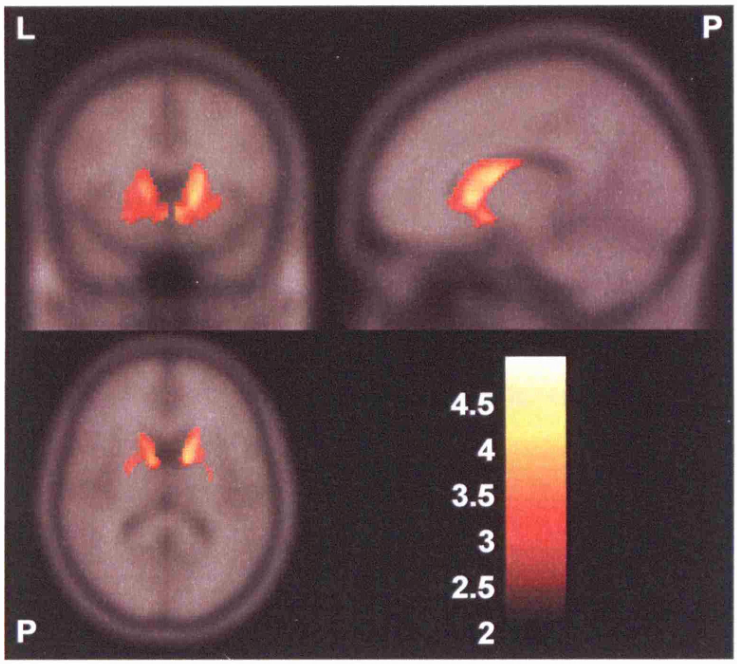


Figure 8-7 Regions in which anger performance was associated with grey matter atrophy in the striatal ROI (coordinates 14, 14, 12mm); q<0.05 FDR, smoothness 9.5, 10.3, 9.6mm, 54.7 resels

8.3.2.5 Fear

In the striatal ROI worse fear recognition in HD gene carriers was associated with atrophy of the putamen and caudate head bilaterally. Fear recognition was also associated with small regions of atrophy in the right insula (but not the left), and with the left middle frontal gyrus and right inferior frontal gyrus in the OFC ROI (Figure 8-8). Better fear recognition in controls, but not HD gene carriers, was associated with increased amygdala volume bilaterally (SPM not shown).

8.3.2.6 Other contrasts

For all the emotions tested there were no significant interactions between slope and group (i.e. regions in which the slope of the association in the HD gene carriers was significantly greater than that in the control group). We also investigated the reverse associations, i.e. looking for regions in which increased brain volume was associated with worse emotion recognition performance in the HD group. There was no evidence of any such regions for any of the emotions tested.

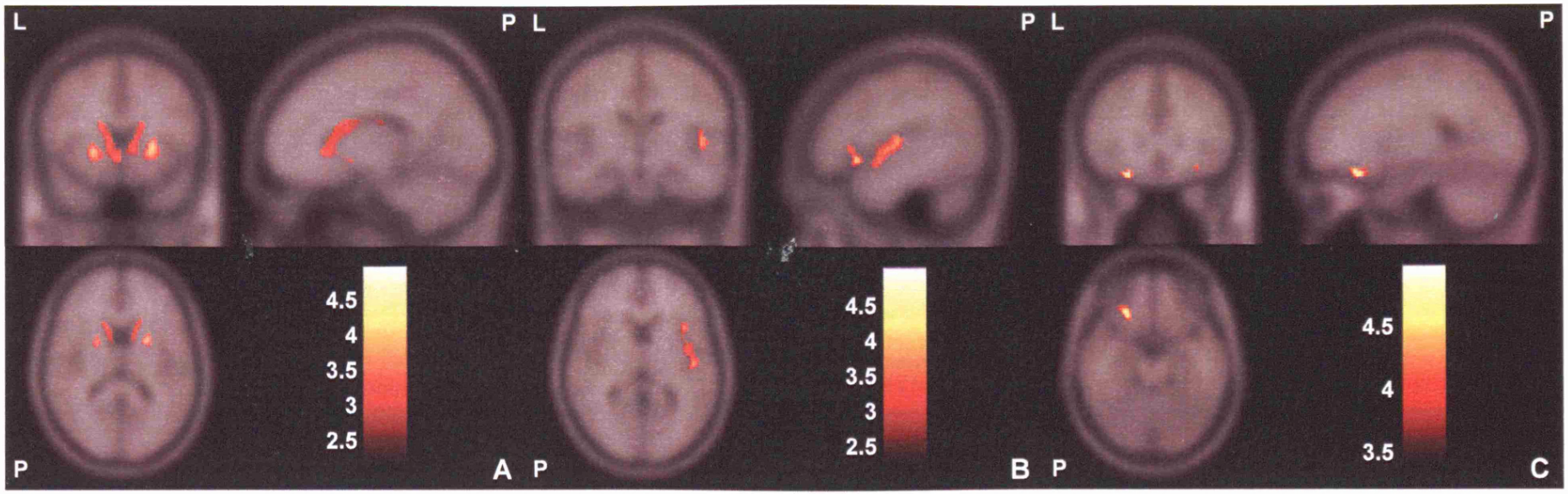


Figure 8-8 Regions in which fear performance was associated with grey matter atrophy in A) striatal ROI (14, 14 12mm) smoothness 9.5, 10.2, 9.6mm, 54.9 resels; B) right insula ROI (47, -15, 6mm), smoothness 9.5, 10.2, 96.mm 28.3 resels; C) orbitofrontal ROI (-25, 29, -19mm) smoothness 9.5, 10.2, 9.6mm, 67.0 resels; all q<0.05 FDR

Table 8-4 Summary table of anatomical correlates for each emotion in each ROI for the HD gene carriers

	Striatum	Insula (L or R)	Orbitofrontal cortex	Amygdala (L and R)
Surprise	Caudate body bilaterally	-	-	-
Disgust	Left caudate head and caudate body bilaterally	-	-	-
Anger	Head and body of caudate bilaterally, right putamen	-	-	-
Fear	Putamen and caudate head bilaterally	Right anterior and posterior	Left middle frontal gyrus, right inferior frontal gyrus	-

8.4 DISCUSSION

The work presented here demonstrates impaired recognition of negative facial emotions in a well-characterised cohort of 40 subjects with early HD. The early stage non-depressed HD subjects were significantly worse than healthy subjects at recognising facial expressions of anger, disgust, fear and surprise, and significantly worse at recognising anger than disgust or fear. Although there was no evidence of significant deficits in individuals with premanifest HD, for all emotions other than sadness there were small effects in the same direction.

Despite the focus on disgust recognition, the original study to report a disproportionate impairment for this emotion also found that recognition of angry faces was impaired (Sprengelmeyer *et al.* 1996) as have others (Milders *et al.* 2003; Wang *et al.* 2003; Montagne *et al.* 2006; Johnson *et al.* 2007), and a similar deficit has recently been demonstrated for recognition of angry body language (de Gelder *et al.* 2008).

There was no evidence of impaired sadness recognition in these HD subjects: although some studies do report this (Sprengelmeyer *et al.* 1996; Milders *et al.* 2003; Johnson *et al.* 2007), in most cases the impairment is slight relative to other negative emotions. This suggests that particular negative emotions may be intrinsically more difficult to recognise, perhaps because they impose greater demands on perceptual processing resources (for example, the need to disambiguate the facial expressions of disgust and anger, which share a number of configurational features (e.g., Simon *et al.* 2007)). In common with most other studies of HD this cohort was not significantly impaired at recognising happiness; this was the least difficult emotion for the controls to recognise and,

as the only positive emotion in the test, is easily contrasted with the other five. This is likely to reflect in part the relative paucity of universally agreed subcategories within the general category of 'positive' emotion, in contrast to the well-established subcategories within the field of 'negative' emotion. The apparently inferior performance for recognition of happiness shown by individuals with premanifest HD relative to healthy controls is difficult to interpret as all groups performed at or near ceiling for this emotion. In fact, given that controls scored entirely at ceiling it is unlikely that this represents a clinically meaningful difference. The lack of a statistically significant difference in the early HD group is probably due to the increased variance in that group.

Impaired recognition of facial emotions in HD was not attributable simply to a perceptual processing defect, since effects persisted after adjusting for perceptual factors. However this does not rule out the possibility that an abnormality of facial feature analysis contributed to the emotion recognition defect. I noted that some subjects who scored poorly on the test seemed to attend only to features of the faces in isolation; for example they would comment on the eyes, and then the mouth, and then justify their final decision based on just one rather than the features as a whole. While this strategy works for some emotions (e.g. all and only the happy stimuli have a smiling mouth) it does not for others (e.g. the angry and disgusted stimuli all have lowered eyebrows and a range of mouth positions). Differences in patterns of visual scanning with ageing have been associated with the (in)ability to recognise certain emotions (e.g., Wong *et al.* 2005). It would be interesting therefore to examine the pattern of subjects' eye movements as they scan the images, since eye movements are known to be impaired in premanifest and early HD (e.g., Golding *et al.* 2006) or to see if

performance could be improved by explicit instructions to attend to specific facial features, as has been shown with amygdala damage (Adolphs *et al.* 2005). However, it is likely that there is a higher-level, supramodal failure of emotion recognition even taking into account specific strategies of facial feature analysis; this would follow from evidence for cross-modal emotion recognition deficits in both HD and other conditions (e.g., Scott *et al.* 1997; Calder *et al.* 2000).

Recognition deficits for anger, disgust, fear and surprise correlated with grey matter loss in a number of brain regions known to be involved in HD. Impaired recognition of anger, the most severe deficit in this cohort, was associated with volume loss in the ventral putamen and other striatal regions. Damage to the ventral putamen has been shown previously to disrupt anger recognition (Calder *et al.* 1996; Scott *et al.* 1997; Calder *et al.* 2004). In HD striatal atrophy progresses in a dorsal to ventral direction (Vonsattel *et al.* 1985), which may mean there was sufficient inter-subject variability in involvement of the ventral striatum in this cohort to detect the correlation with anger recognition performance. Impaired recognition of disgust was also associated with atrophy of the striatum, but not the insula or globus pallidus; these latter regions have been implicated in disgust recognition in other studies, including premanifest HD (Krolak-Salmon *et al.* 2003; Murphy *et al.* 2003; Kipps *et al.* 2007). Comparison with the healthy control group confirmed that bilateral insula atrophy was present in this HD cohort (see chapter 7), suggesting that any association between disgust recognition performance and insula volume was weak, or perhaps that uniform involvement of the insula precluded correlation with performance in the HD group. Impaired fear recognition was also associated with striatal volume loss, although the effect here was weaker than for anger and disgust. At least one

fMRI study has found a similar relationship between fear and the striatum (Phillips *et al.* 1998). The fear recognition deficit in this cohort was associated with additional atrophy of the right insula, and lateral orbitofrontal cortex. The insula has been implicated in the prediction of aversive stimuli and the behavioural relevance of stimuli, (Buchel *et al.* 1998; Hennenlotter *et al.* 2004; Simmons *et al.* 2004), while the orbitofrontal cortex is involved in modulating behaviour in response to emotive stimuli in general (Sprengelmeyer *et al.* 1998; Blair *et al.* 1999; Gorno-Tempini *et al.* 2001). There was no evidence that fear recognition was related to amygdala volume in the HD group. This association was present in controls, although this is hard to interpret without further investigation. This may reflect the distinction between categorisation of emotion in others as in the task used here, versus subjective feeling states, for which the amygdala may be critical (e.g., Lang *et al.* 2000b; Calder *et al.* 2001; Skuse *et al.* 2003). The association between impaired surprise recognition and caudate atrophy in the present study was relatively weak and further work is needed to confirm this. There is little evidence for a specific neural substrate for surprise recognition (Murphy *et al.* 2003).

These associations between brain volume and emotion recognition performance were found in the HD gene-carrying cohort, and not in controls, but the difference between the associations in the two groups was not significant, meaning that disease-specificity of these effects was not shown. However the findings are suggestive of such an effect, and most likely reflect a lack of power which could be addressed with a larger sample size. Indeed some studies infer disease-specificity from associations seen in the HD group alone (Douaud *et al.* 2006). It may also be useful in future work to model volume as a function of all

the emotions as one could then investigate which regions were associated with one emotion and not others. However to have adequate power such a model would probably require a larger dataset than that presented here and as such was beyond the scope of this thesis.

Whilst these findings are in general agreement with other studies showing impaired recognition of negative emotions in HD (Sprengelmeyer *et al.* 1996; Gray *et al.* 1997; Milders *et al.* 2003; Johnson *et al.* 2007; Kipps *et al.* 2007), the pattern of this impairment differs between studies. Whereas Grey *et al.* (1997) and Sprengelmeyer *et al.* (1996) described selective deficits in disgust recognition, Milders *et al.* (2003) found fear to be most severely affected, while anger was most severely affected in the present study. The data presented here support the idea that damage to a distributed network of brain regions underpins a broad emotion recognition deficit in early HD; the symptomatic cohort was at a very early stage of disease and yet was already impaired at recognising a range of negative emotions. The common association of these deficits with striatal volume loss suggests that the striatum might play a generic role in the processing of negative emotions. However, it is likely that emotion recognition deficits in HD arise from damage involving a cortico – basal ganglia – thalamic network, rather than focal damage to a specific area. The ventral part of this network, which is thought to process emotional or reward information, receives input from orbitomedial prefrontal regions and insula and outputs predominantly from the thalamus via the basal ganglia (Parent and Hazrati 1995; Haber 2003). These regions are involved even in very early or premanifest HD (e.g., Thieben *et al.* 2002; Rosas *et al.* 2005). The present findings imply that emotion recognition deficits in HD may appear selective but represent 'snapshots' of a more generic

structure–function relationship that is distributed in the brain and perhaps also evolving in time. Involvement of the network may not be uniform for all emotions: rather, the difference between emotions may emerge as a relatively greater or lesser involvement or correlated involvement of particular components of the shared network.

It should also be noted that use of ROIs in this study means that it cannot be concluded that brain regions outside those investigated here are not involved in emotion recognition in HD. Although no statistically significant effects were found in the whole-brain analysis the effect maps showed that there were positive relationships between volume and score in the temporal lobes (for surprise and fear recognition in particular) and the posterior cortex (fear), as well as in the striatum, insula and orbitofrontal cortex. This is something that merits further investigation with both structural and functional imaging.

8.4.1 Methodological considerations

As with findings throughout the HD literature, the variability amongst studies of emotion recognition in HD might reflect a number of different factors, including the characteristics of the HD population under investigation, the test used, and the choice of control group. Different study cohorts may represent different disease stages with varying burdens and patterns of brain atrophy. For example, it is likely that in the earliest stages of the disease striatal atrophy is more severe than insula. It may also be that linguistic and cultural differences contributed to the diverse findings, as early studies were undertaken in Germany in contrast to the later studies which have been predominantly in English-speaking countries.

However, in addition the three main tools for assessing facial emotion recognition in the literature – the "emotion hexagon" which uses morphs of a single subject; a set of 60 Ekman faces; and the set of 24 Ekman faces used here and by Gray *et al.* (1997) – are not of equivalent difficulty (based on the mean percent correct achieved by Ekman's normative sample). In addition, the emotion hexagon is not a simple facial emotion recognition task. Morphs of a single subject ("JJ") were constructed by blending different proportions of two emotions, placing each next to one it was most likely to be confused with (Calder *et al.* 1996) although in fact, in order to fit all six canonical emotions into the hexagon this is not always true for each pair of emotions. Blocks tend to be repeated in testing which might inflate differences between controls and HD subjects if controls benefited from learning over the earlier blocks whilst HD subjects did not.

Investigators have also used widely varying criteria for selecting controls, ranging from normal healthy adults, to at-risk subjects who are undergoing genetic testing, to gene-negative and HD family members, and these groups are unlikely to be equivalent. Impaired anger recognition has been shown in at-risk controls who subsequently turn out to be gene negative (Gray *et al.* 1997; Sprengelmeyer *et al.* 2005), and there is also evidence that fear and anger recognition might deteriorate slightly with age, whilst disgust recognition is relatively spared (Calder *et al.* 2003). The original finding of impaired disgust recognition in HD used controls who were on average five years older than HD subjects, and did not control for the effects of age (Sprengelmeyer *et al.* 1996). The controls in the current study found fear most difficult to recognise, followed by sadness and anger, and their performance on these emotions was significantly

worse than that of Ekman's (presumably younger) normative college student sample (Table 8-3); however the overall rank ordering of scores was similar.

8.5 CONCLUSION

These findings suggest a widespread impairment of facial emotion recognition in HD. Anger, disgust, fear and surprise recognition were all affected to a variable degree early in the disease process, and these deficits were associated with reduced grey matter volume in a network of cortical and subcortical regions. Striatal damage was associated with impaired recognition of different negative emotions. It is likely that these brain regions constitute a distributed functional network that mediates emotion recognition: this network may have a generic neurobiological role in processing threat to self, whether direct external (anger), indirect external (fear) or internal (disgust), and is therefore predicted to underpin a number of aspects of human social behaviour and behavioural derangements in disease. This hypothesis should motivate further work, including longitudinal structural and functional brain imaging, in order to clarify when deficits become manifest and how their brain substrate evolves over time. From the clinical perspective, emotion recognition deficits may be useful in determining the onset of cognitive involvement and in tracking disease progression in HD. Anecdotally, many of the carers in this study described episodes during which their HD-affected partner failed to recognise that they were in some way upset, and the distress that this caused. A better understanding of the pattern of deficits might also therefore have implications for anticipating and managing the often devastating social difficulties faced by people with HD and their carers.

9 DETECTING COGNITIVE DECLINE IN PREMANIFEST AND EARLY HUNTINGTON'S DISEASE

9.1 INTRODUCTION

As mentioned in section 1.4.1, tracking change in cognition in HD is difficult, because of the many non-disease-related factors that can affect performance, and the slow progression of the disease. There have been conflicting findings as to which domains show decline in early HD and few studies have been able to detect cognitive decline in premanifest subjects using standard cognitive measures.

In this study a number of factors which might contribute to change in cognitive performance were considered and where possible, controlled for. Subjects were all assessed by the same investigator, using alternative test forms at the follow-up assessment where they were available. Testing was done in similar surroundings although it was not always possible to assess subjects in the same room or at the same time of day at both assessments.

Estimated premorbid IQ at baseline was adjusted for, since it could independently affect cognitive decline (Snowden *et al.* 2001; Ward *et al.* 2006). Age was also included as a covariate. Although in previous studies the effect of age on rate of decline was negligible, the current cohort spanned a wide range of ages and preliminary analysis showed that age had a small but sometimes statistically significant effect on the rate of decline in controls, particularly for timed tests.

This chapter describes changes over one year in cognition, whilst the following chapters look at changes in brain volume and the associations between the two. The aims of this chapter were to investigate in which domains cognitive decline could be detected, over a relatively short time period, and how change was related to CAG repeat length and other baseline demographic variables.

9.2 METHODS

9.2.1 Subjects and assessments

Subjects and assessments were the same as in chapter 4 (sections 4.2.1 and 4.2.2). All subjects from the baseline assessment were invited to return for repeat cognitive, clinical and MRI assessments approximately one year later. As described in chapter 4, where alternative test versions were available half the subjects did each form at each timepoint. All subjects who returned for follow-up are reported in this chapter (77 of the 81 assessed at baseline, see section 9.3 for demographic details). The vast majority, 72 subjects, had all three follow-up assessments on the same day or over two consecutive days. Three (two early HD subjects and one control) had the clinical assessment after the other two assessments (five, seven and 18 weeks later); and one early HD subject had the scan and clinical assessment eight days before the cognitive assessment for logistical reasons. One early HD subject completed the cognitive assessment but suffered from anxiety during the scan and had to return nine weeks later to complete the MRI and clinical assessments.

9.2.2 Statistical analysis

9.2.2.1 Group differences

t-tests assuming unequal variance were used to investigate group differences in age, estimated premorbid IQ and CAG repeat length at baseline. A chi^2 test was used to compare gender in each group. Fisher's exact test was used to see if the proportion of right-handed subjects, or the proportion of subjects classified as depressed at baseline or follow-up, differed between groups.

As controls were at ceiling on the UHDRS motor score, independence score and TFC, and premanifest were at ceiling on the independence score and TFC, paired t-tests were used to assess change between visits within each group (i.e., change in premanifest and early HD subjects was not compared with change in controls). Control motor scores did change very slightly over the year and are presented for comparison only.

In order to investigate whether performance differed significantly between cognitive test versions, follow-up score was regressed on test form, adjusting for age and estimated premorbid IQ, for all tests for which an alternative version was available.

As slightly fewer subjects returned for follow-up than were assessed at baseline, linear regression models with robust standard errors were used to investigate group differences in baseline score (as in chapter 4) for the 77 subjects who returned for follow-up.

For cognitive variables a change score was generated from the baseline and follow-up score. Linear regression models were used to compare change scores

between groups, with robust standard errors to allow variances to differ between groups, and controlling for age and estimated premorbid IQ at baseline by including them as covariates. Change scores for the Ekman emotion recognition task were not Normally distributed and so linear regression models were used with a 95% bias-corrected bootstrap confidence interval with 1000 replicates.

Examination of the raw data showed that some subjects scored at ceiling on the HVLT delayed recall and discrimination tasks, and the RMW, at both timepoints, and hence change scores are unlikely to be very sensitive in this cohort. Floor or ceiling effects were not apparent in the other tasks.

Data from the reaction time task were unable to be analysed longitudinally as an unavoidable software change half-way through the follow-up meant that reaction times differed systematically between some subjects.

9.2.2.2 Relationship with CAG repeat length

A linear regression model (with robust standard errors) relating cognitive change to CAG repeat length, group and their interaction was used to investigate whether the relationship between change and CAG repeat length differed between groups. The effects of age and estimated premorbid IQ were controlled for by including them as covariates. If there was no evidence of an interaction then the analysis was repeated without interaction terms, but adjusting for differences due to group, and age and estimated premorbid IQ.

9.2.2.3 Relationship with baseline motor score, disease duration and interaction between CAG repeat length and duration

Separate linear regression models were used to investigate the association between cognitive change and both baseline motor score, and disease duration (at baseline) in the early HD group, again controlling for the effects of age and estimated premorbid IQ at baseline. As duration and age at baseline linearly combine to give age at onset, this model implicitly investigates the effect of age at onset after adjustment for baseline age, as well as the effect of disease duration after adjustment for baseline age.

Finally, in view of the recent finding that both repeat length and disease duration had an interactive effect on rate of decline (Rosenblatt *et al.* 2006), change scores were also regressed on CAG repeat length, duration and their interaction, with age and IQ included as covariates.

9.2.2.4 Relationship between probability of onset and other variables

Probability of motor onset within five years was calculated using the equation given by Langbehn *et al.* (2004) (see section 4.2.3.5 for more details). To assess whether rate of cognitive change was influenced by closeness to motor onset in the premanifest group cognitive change scores were regressed on probability of onset, controlling for the effects of age and IQ by including them as covariates.

9.2.2.5 Comparison of NART and Spot the Word test

In order to compare the stability of these two tests over time, firstly the Spot the Word test was transformed to be on the same scale as the NART (converting a distribution with a mean of 10 and standard deviation of 3 to one with a mean of 100 and a standard deviation of 15). Paired t-tests were used to investigate

whether there was a significant amount of change in either test between baseline and follow-up, in the whole cohort. For each measure a change score was generated (follow-up score - baseline score) and t-tests assuming unequal variance were used to determine whether the amount of change in each test differed between groups.

Pitman's test was used to compare the standard deviation of the change score for each test. A paired t-test was used to investigate whether the amount of change differed significantly between tests.

9.3 RESULTS

Two controls and two premanifest subjects did not return for the follow-up assessment. Demographic data for the remaining 77 subjects are shown in Table 9-1. There were small non-significant differences in gender, handedness and IQ between the groups. There was no evidence of a difference in assessment interval between the groups. The mean age of the premanifest group was eight and a half years lower than that of controls (95% CI 2.5, 14.6 years, p=0.006). The premanifest group had significantly shorter CAG repeat lengths, on average, than the early HD group (mean difference 1.5 repeats, 95% CI 0.3, 2.8, p=0.015). One control (6%), five premanifest (26%) and 12 early HD subjects (33%) were classified as depressed at baseline; the BDI was not available to be administered to four early HD subjects at baseline (see section 4.2.2.1). Five controls (28%), five premanifest (26%) and 13 early HD subjects (33%) were classified as depressed at follow-up and these proportions did not differ significantly between groups at either timepoint (baseline p=0.07, follow-up p=0.90).

Table 9-1 Demographic data for the longitudinal cohort

	Control (N=18)	Premanifest (N=19)	Early HD (N=40)
Gender (M:F)	6:12	9:10	20:20
Age (year)[a]	46.3 (9.7)	37.7 (7.6)	48.5 (9.6)
Estimated premorbid IQ	106.8 (11.5)	104.2 (9.1)	105.3 (13.0)
Handedness (R:L)	17:1	18:1	36:4
CAG repeat length[b]	NA	42.2 (1.8), range 40 – 45	43.7 (2.4), range 40 – 50
Predicted years to onset[c]	NA	17.8 (7.4), range 9 – 35	NA
Disease duration (year)	NA	NA	4.1 (2.6)
UHDRS Motor[d]	1.1 (0.9)	3.6 (4.3)	28.9 (12.6)
UHDRS Independence	100 (0)	100 (0)	90.4 (9.6)
UHDRS TFC	13 (0)	13 (0)	10.9 (1.8)
BDI	5.4 (3.3)	6.2 (5.9)	9.3 (8.7)
Assessment interval (year)	1.0 (0.1)	1.0 (0.1)	1.0 (0.1)

NA: not applicable

Data are mean (SD) with the exception of gender and handedness; handedness was taken as the hand used to write with; UHDRS: motor is out of 124, higher score = more severely impaired; independence is scored as a percentage, higher score = better function; Total Functional Capacity is out of 13, higher score = better function.

a PM<Control ($p=0.006$); PM<HD ($p<0.0001$)

b PM<HD ($p=0.015$)

c Calculated from age using the equation from Langbehn *et al.* (2004), with onset defined as a 60% chance of showing motor signs

d HD<Control, HD<PM (both $p<0.0001$)

After correcting for age and IQ there were no statistically significant differences between scores on alternative test versions, except for the HMGT at baseline in which subjects performed worse on form B than form A. As half of each group did each form this is unlikely to bias between-group comparisons but form was included as a covariate in all analyses of the HMGT in order to control for this. One control had no motor assessment at follow-up. Three HD subjects failed to complete TMT B at baseline, and four failed to complete it at follow-up so change scores for TMT B are based on data from 72 subjects. In all cases subjects made errors on the alternating sequencing and struggled to work out the correct number (or letter) from which to continue, and if this carried on for more than four minutes the test was stopped.

9.3.1 Clinical measures

Change in UHDRS motor score, independence score, and TFC for the groups is shown in Table 9-2. On average the early HD group declined in all three measures over one year (see table for details). Change between baseline and follow-up motor score in the control and premanifest groups was not statistically significant although was approaching this in the latter group (p=0.07). On average all subjects scored slightly higher on the BDI at follow-up compared with baseline although this difference was not statistically significant (mean change 0.7 points, 95% CI -0.5, 2.0 points, p=0.25), and there was no evidence that the amount of change differed between the groups.

Table 9-2 Mean (SD) UHDRS scores in the patient groups at baseline and 12-month assessments, with change (95% confidence intervals)

	Control (N=17)		
	Baseline	12 months	Change
Motor	1.1 (0.9)	2.1 (2.4)	1.0 (-0.2, 2.2) p=0.10

	Premanifest (N=19)		
	Baseline	12 months	Change
Motor	3.6 (4.3)	5.2 (4.4)	1.6 (-0.2, 3.4) p=0.07

	Early HD (N=40)		
	Baseline	12 months	Change
Motor	28.9 (12.6)	34.3 (12.9)	5.4 (3.0, 7.7) p<0.0001
Independence score	90.4 (9.6)	88.6 (9.3)	-1.8 (-3.1, -0.4) p=0.0114
TFC	10.9 (1.8)	10.4 (1.9)	-0.4 (-0.7, -0.1) p=0.0112

9.3.2 Cognitive performance

9.3.2.1 Group differences at baseline

In general group differences at baseline mirrored those seen in the whole cohort (N=81, see chapter 4) and will be summarised here for brevity. The early HD group was impaired, relative to controls, at exactly the same tests, namely all tests of psychomotor skills, reaction time, executive function, immediate, free recall and recognition memory, digit span, Benton facial recognition, and Ekman facial emotion recognition. The premanifest group was impaired at returning to press the central blue button in the RT task. In addition, in an effect not seen at baseline, this group was impaired at HVLT discrimination ($p=0.035$). The two subjects who withdrew were estimated to be 19 and 23 years from onset, i.e. even within the far-from-onset cohort presented here they had greater than average time to onset. Their withdrawal meant that the remaining 19 were slightly closer to onset, on average, and which may explain the finding of this deficit. As with the full cohort, neither group was impaired at naming or visuoperceptual skills.

9.3.2.2 Group differences in cognitive decline

In general the early HD group tended to decline slightly at the majority of tests, with the exception of the GNT, VOSP, emotion recognition, Stroop interference, cancellation, and the two parts of the HVLT on which many were performing at ceiling. In contrast the control group tended to show slight improvements, with the exception of Stroop word, category fluency and TMT A (all of which are timed) and most subtests of the HVLT (Table 9-3).

After adjusting for estimated premorbid IQ and age at baseline the mean change over time in the early HD group was significantly different to that in controls for the UHDRS cognitive score (p=0.001), RMF (p=0.047) RMW (p=0.02) and TMT B (p=0.01), with the HD group declining whilst the controls improved slightly. Although some early HD subjects had scored at ceiling on the RMW at baseline none had at follow-up, suggesting that the test was sensitive at detecting decline. However six controls were at ceiling at follow-up, so relative to the control group this test may still have over- or under-estimated group differences (Table 9-3).

Change in the premanifest group was similar to that in controls, in that performance in most tests improved slightly on average, with the exception of Stroop word and interference, cancellation, and two subtests of the HVLT (Table 9-4).

Both groups showed statistically significantly different change to controls at phonemic fluency (early HD vs. control p=0.003, premanifest vs. control p=0.05) although the difference in this case was driven by a large improvement in controls (Figure 9-1). There were no other statistically significant changes in either HD group relative to controls (Table 9-3, Table 9-4).

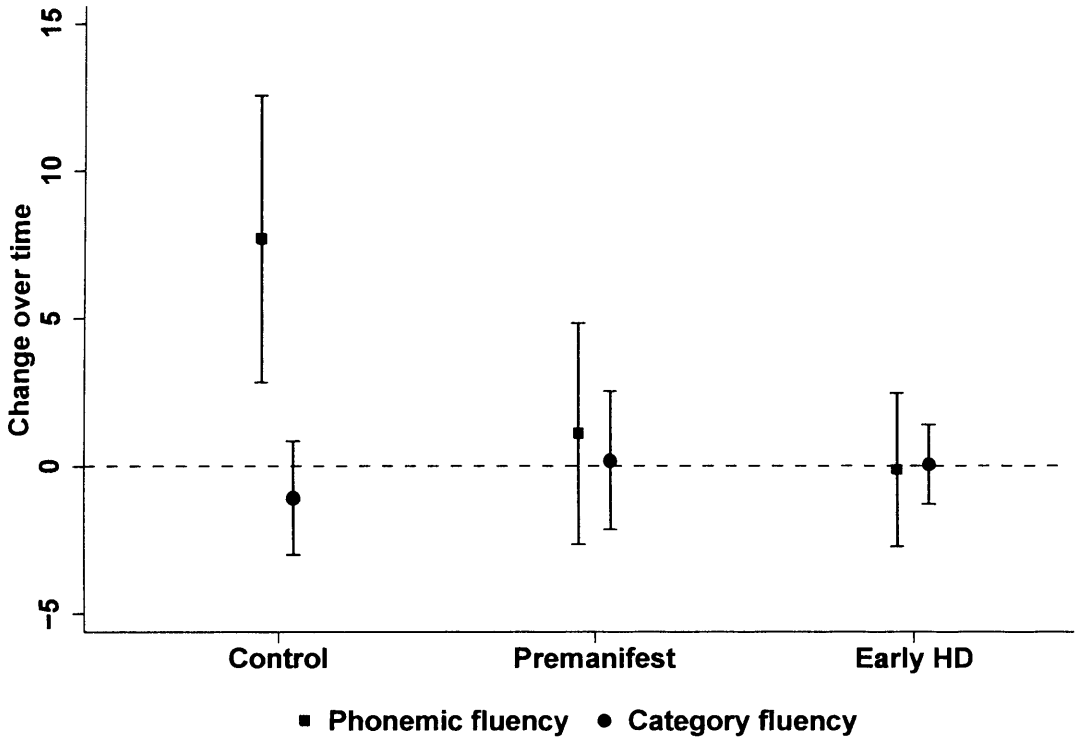


Figure 9-1 Change in phonemic and category fluency scores in each group, adjusted for age and estimated premorbid IQ at baseline, with 95% confidence intervals

Negative change indicates decline over time.

Table 9-3 Mean (SD) baseline and change in cognitive scores for controls and early HD subjects, with differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ at baseline

	Control (N=18)		Early HD (N=40)		Difference in change (Early HD - Control)	
Test	Baseline	Change	Baseline	Change	Crude	Adjusted
UHDRS cognitive (n)	335.6 (35.6)	10.2 (16.7)	231.5 (53.6)	-7.4 (19.3)	-17.6 (-27.5, -7.7) p=0.001	-17.6 (-27.3, -7.8) p=0.001
Stroop colour (n)	80.7 (9.6)	1.7 (6.2)	57.0 (12.7)	-1.5 (7.5)	-3.1 (-6.9, 0.6) p=0.10	-2.8 (-6.4, 0.9) p=0.14
Stroop word (n)	107.3 (13.6)	-2.7 (8.4)	75.4 (21.2)	-5.0 (10.0)	-2.3 (-7.4, 2.7) p=0.37	-2.7 (-7.7, 2.2) p=0.27
Stroop interference (n)	46.4 (7.8)	1.7 (3.6)	32.8 (9.2)	0.4 (5.1)	-1.3 (-3.7, 1.0) p=0.26	-1.4 (-3.9, 1.0) p=0.25
SDMT (n)	55.8 (8.3)	1.8 (5.5)	33.3 (10.3)	-0.9 (4.8)	-2.8 (-5.7, 0.2) p=0.07	-2.8 (-5.9, 0.2) p=0.07
Phonemic fluency (n)	45.4 (9.1)	7.6 (9.8)	33.0 (11.3)	-0.5 (8.2)	-8.1 (-13.3, -2.8) p=0.003	-7.8 (-13.0, -2.7) p=0.003
Category fluency (n)	24.1 (4.1)	-1.1 (4.0)	17.9 (5.6)	0.0 (4.2)	1.1 (-1.2, 3.3) p=0.37	1.1 (-1.0, 3.3) p=0.31
HMGT (n)	12.1 (2.3)	0.4 (3.1)	10.25 (2.5)	-0.4 (2.3)	-1.0 (-2.2, 0.3) p=0.15	-1.0 (-2.3, 0.4) p=0.16
TMT A (sec)	21.1 (6.6)	-0.8 (7.3)	37.6 (11.5)	-1.5 (13.3)	-0.7 (-6.1, 4.7) p=0.79	-0.9 (-6.1, 4.2) p=0.72
TMT B (sec)	64.7 (42.7)	10.1 (30.0)	119.6 (55.6)	-15.3 (38.4)	-25.3 (-44.5, -6.2) p=0.01	-26.0 (-45.6, -6.5) p=0.01

	Control (N=18)		Early HD (N=40)		Difference in change (Early HD - Control)	
Test	Baseline	Change	Baseline	Change	Crude	Adjusted
Cancelling As (sec)	16.8 (4.0)	0.7 (1.3)	28.1 (8.5)	0.3 (5.3)	-0.4 (-2.2, 1.4) p=0.65	-0.5 (-2.3, 1.2) p=0.56
HVLT imm. (/36)	26.1 (6.1)	2.1 (5.6)	19.7 (5.2)	-0.5 (5.7)	-2.6 (-5.7, 0.6) p=0.11	-2.6 (-6.0, 0.7) p=0.13
HVLT del. (/12)	9.9 (1.8)	-0.1 (1.7)	6.4 (2.4)	0 (1.9)	0.1 (-0.9, 1.0) p=0.91	0.1 (-0.9, 1.1) p=0.90
HVLT % recall	97.0 (11.5)	-7.8 (15.9)	80.2 (24.0)	-0.8 (25.4)	7.0 (-3.9, 18.0) p=0.21	6.8 (-4.8, 18.3) p=0.25
HVLT discrim. (/12)	11.2 (1.0)	-0.3 (1.3)	8.8 (2.3)	0.1 (2.2)	0.4 (-0.5, 1.3) p=0.41	0.3 (-0.6, 1.3) p=0.50
Digits forwards (/21)	13.1 (3.8)	0.8 (2.2)	10.5 (3.2)	-0.1 (1.9)	-0.9 (-2.1, 0.3) p=0.14	-0.9 (-2.1, 0.3) p=0.14
Digits backwards (/21)	12.1 (4.1)	0.4 (2.7)	8.7 (3.8)	-0.6 (3.1)	-1.0 (-2.6, 0.6) p=0.22	-1.1 (-2.7, 0.5) p=0.19
RMF (/50)	42.4 (4.0)	1.8 (3.5)	34.8 (7.0)	-0.7 (5.3)	-2.5 (-4.8, -0.2) p=0.036	-2.5 (-4.9, -0.03) p=0.047
RMW (/50)	48.0 (2.5)	0.1 (1.7)	43.4 (4.7)	-1.5 (3.8)	-1.4 (-3.0, -0.1) p=0.031	-1.7 (-3.2, -0.3) p=0.02
GNT (/30)	21.6 (3.8)	0.4 (1.7)	20.7 (4.7)	0.3 (1.9)	-0.1 (-1.2, 0.8) p=0.70	-0.1 (-1.1, 0.8) p=0.78
VOSP (/15)	9.4 (2.4)	0.8 (1.7)	8.6 (2.3)	0.6 (1.3)	-0.2 (-1.1, 0.7) p=0.61	-0.2 (-1.1, 0.7) p=0.61
Benton (/54)	47.7 (4.0)	0.9 (4.1)	43.5 (4.7)	-0.9 (4.2)	-1.9 (-4.2, 0.4) p=0.11	

	Control (N=18)		Early HD (N=40)		Difference in change (Early HD - Control)	
Test	Baseline	Change	Baseline	Change	Crude	Adjusted
Ekman happiness (/4)	4.0 (0.0)	0.0 (0.0)	3.9 (0.3)	0.03 (0.5)	0.03 (-0.1, 0.2) p>0.05	0.02 (-0.1, 0.2) p>0.05
Ekman sadness (/4)	3.4 (0.9)	-0.1 (0.6)	3.3 (0.8)	-0.1 (0.7)	0.01 (-0.4, 0.4) p>0.05	-0.02 (-0.4, 0.4) p>0.05
Ekman surprise (/4)	3.8 (0.4)	-0.1 (0.5)	3.4 (0.6)	-0.03 (1.1)	0.1 (-0.4, 0.5) p>0.05	0.1 (-0.4, 0.5) p>0.05
Ekman disgust (/4)	3.7 (0.5)	-0.1 (0.7)	2.7 (1.2)	0.1 (1.0)	0.2 (-0.2, 0.6) p>0.05	0.2 (-0.2, 0.6) p>0.05
Ekman anger (/4)	3.6 (0.6)	0.0 (0.6)	2.0 (1.3)	-0.1 (1.0)	-0.1 (-0.5, 0.4) p>0.05	-0.1 (-0.5, 0.4) p>0.05
Ekman fear (/4)	2.9 (0.8)	0.4 (0.7)	1.7 (1.3)	0.1 (1.4)	-0.3 (-0.9, 0.2) p>0.05	-0.4 (-1.0, 0.2) p>0.05
Ekman total (/24)	21.4 (1.6)	0.2 (1.2)	16.8 (3.3)	0.1 (2.4)	-0.1 (-1.0, 0.9) p>0.05	-0.2 (-1.1, 0.9) p>0.05

Negative change indicates a decrease in score or in speed (longer time taken in timed tasks); negative difference in change indicates that early HD declined more than controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models; Benton score is already age- and education adjusted and hence only crude difference is reported; where 95% bootstrapped CIs were used precise p values were not obtained.

Table 9-4 Mean (SD) baseline and change in cognitive scores for controls and premanifest subjects, with differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ at baseline

	Control (N=18)		Premanifest (N=19)		Difference in change (Premanifest – Control)	
Test	Baseline	Change	Baseline	Change	Crude	Adjusted
UHDRS cognitive	335.6 (35.6)	10.2 (16.7)	326.4 (51.1)	2.5 (20.5)	-7.6 (-20.1, 4.8) p=0.23	-6.9 (-20.4, 6.6) p=0.31
Stroop colour	80.7 (9.6)	1.7 (6.2)	78.5 (9.7)	2.3 (6.6)	0.6 (-4.0, 5.2) p=0.80	0.1 (-4.1, 4.4) p=0.94
Stroop word	107.3 (13.6)	-2.7 (8.4)	106.6 (18.2)	-2.6 (6.7)	0.1 (-5.8, 5.9) p=0.98	1.4 (-4.1, 7.0) p=0.62
Stroop interference	46.4 (7.8)	1.7 (3.6)	47.7 (9.9)	-1.2 (7.5)	-2.9 (-6.7, 0.9) p=0.14	-2.5 (-6.5, 1.4) p=0.20
SDMT	55.8 (8.3)	1.8 (5.5)	53.9 (10.5)	2.2 (4.8)	0.3 (-2.9, 3.6) p=0.84	0.7 (-2.9, 4.3) p=0.69
Phonemic fluency	45.4 (9.1)	7.6 (9.8)	39.6 (14.0)	1.8 (7.7)	-5.8 (-11.3, -0.2) p=0.042	-6.6 (-13.2, -0.002) p=0.05
Category fluency	24.1 (4.1)	-1.1 (4.0)	24.2 (7.2)	0.3 (4.9)	1.4 (-1.5, 4.2) p=0.34	1.3 (-1.9, 4.5) p=0.43
HMGT	12.1 (2.3)	0.4 (3.1)	11.7 (2.8)	0.3 (2.9)	-0.3 (-1.7, 1.2) p=0.72	-0.2 (-1.9, 1.5) p=0.84
TMT A (sec)	21.1 (6.6)	-0.8 (7.3)	22.5 (5.2)	0.9 (5.5)	1.7 (-2.5, 5.9) p=0.43	1.4 (-3.4, 6.1) p=0.57
TMT B (sec)	64.7 (42.7)	10.1 (30.0)	61.8 (25.9)	4.6 (24.4)	-5.5 (-23.3, 12.4) p=0.54	4.2 (-12.2, 20.6) p=0.61

	Control (N=18)		Premanifest (N=19)		Difference in change (Premanifest – Control)	
Test	Baseline	Change	Baseline	Change	Crude	Adjusted
Cancelling As (sec)	16.8 (4.0)	0.7 (1.3)	18.5 (4.7)	-0.4 (4.5)	-1.1 (-3.2, 1.0) p=0.31	-0.6 (-3.3, 2.1) p=0.66
HVLT imm. (/36)	26.1 (6.1)	2.1 (5.6)	26.7 (5.1)	-0.3 (3.2)	-2.3 (-5.3, 0.7) p=0.13	-1.8 (-5.0, 1.5) p=0.28
HVLT del. (/12)	9.9 (1.8)	-0.1 (1.7)	9.3 (2.3)	0.3 (2.1)	0.4 (-0.9, 1.6) p=0.56	0.5 (-0.8, 1.9) p=0.42
HVLT % recall	97.0 (11.5)	-7.8 (15.9)	90.6 (15.6)	0.7 (19.1)	8.5 (-2.9, 19.9) p=0.14	8.4 (-4.1, 20.8) p=0.18
HVLT discrim. (/12)	11.2 (1.0)	-0.3 (1.3)	10.8 (1.2)	-0.1 (1.0)	0.3 (-0.5, 1.0) p=0.45	0.4 (-0.4, 1.2) p=0.30
Digits forwards (/21)	13.1 (3.8)	0.8 (2.2)	13.3 (3.3)	0.3 (2.5)	-0.5 (-1.9, 0.9) p=0.46	-0.5 (-2.2, 1.3) p=0.58
Digits backwards (/21)	12.1 (4.1)	0.4 (2.7)	11.3 (4.1)	0.1 (2.3)	-0.4 (-2.3, 1.5) p=0.68	-0.1 (-1.8, 1.6) p=0.91
RMF (/50)	42.4 (4.0)	1.8 (3.5)	41.1 (5.3)	0.6 (3.7)	-1.2 (-4.2, 1.8) p=0.43	-1.1 (-3.7, 1.5) p=0.40
RMW (/50)	48.0 (2.5)	0.1 (1.7)	48.1 (2.0)	0.6 (1.9)	0.5 (-0.7, 1.6) p=0.43	0.8 (-0.6, 2.2) p=0.26
GNT (/30)	21.6 (3.8)	0.4 (1.7)	22.4 (3.9)	0.4 (1.8)	-0.1 (-1.3, 1.1) p=0.90	-0.4 (-1.6, 0.7) p=0.46
VOSP (/15)	9.4 (2.4)	0.8 (1.7)	8.6 (3.3)	0.9 (2.4)	0.1 (-1.2, 1.4) p=0.86	0.1 (-1.3, 1.6) p=0.85
Benton (/54)	47.7 (4.0)	0.9 (4.1)	47.7 (3.8)	0.9 (3.2)	-0.05 (-2.6, 2.5) p=0.97	

	Control (N=18)		Premanifest (N=19)		Difference in change (Premanifest – Control)	
Test	Baseline	Change	Baseline	Change	Crude	Adjusted
Ekman happiness (/4)	4.0 (0.0)	0.0 (0.0)	3.9 (0.2)	0.1 (0.2)	0.1 (-0.0, 0.2) p>0.05	0.1 (-0.02, 0.3) p>0.05
Ekman sadness (/4)	3.4 (0.9)	-0.1 (0.6)	3.7 (0.6)	0.2 (0.8)	0.2 (-0.2, 0.7) p>0.05	0.3 (-0.1, 0.9) p>0.05
Ekman surprise (/4)	3.8 (0.4)	-0.1 (0.5)	3.7 (0.6)	0.1 (0.6)	0.2 (-0.2, 0.5) p>0.05	0.2 (-0.1, 0.7) p>0.05
Ekman disgust (/4)	3.7 (0.5)	-0.1 (0.7)	3.3 (0.9)	0.1 (1.0)	0.2 (-0.4, 0.7) p>0.05	0.2 (-0.4, 0.6) p>0.05
Ekman anger (/4)	3.6 (0.6)	0.0 (0.6)	3.5 (1.0)	0.2 (0.7)	0.2 (-0.2, 0.6) p>0.05	0.2 (-0.2, 0.7) p>0.05
Ekman fear (/4)	2.9 (0.8)	0.4 (0.7)	2.5 (1.3)	0.3 (1.1)	-0.2 (-0.8, 0.4) p>0.05	-0.2 (-0.8, 0.4) p>0.05
Ekman total (/24)	21.4 (1.6)	0.2 (1.2)	20.6 (2.4)	0.8 (1.9)	0.6 (-0.3, 1.7) p>0.05	0.9 (-0.2, 2.0) p>0.05

Negative change indicates a decrease in score or in speed (longer time taken in timed tasks); negative difference in change indicates that premanifest declined more than controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models; Benton score is already age- and education adjusted and hence only crude difference is reported; where 95% bootstrapped CIs were used precise p values were not obtained.

9.3.2.3 Relationship with CAG repeat length

In general there was very little evidence of any association between CAG repeat length and the amount of change on the UHDRS, BDI or cognitive scores. In the premanifest group higher CAG repeat length was associated with greater decline in digit span forwards such that an increase of one triplet was associated with a 0.9 point decline in digit span forwards (95% CI -1.5, -0.4 points, p=0.001), but this relationship was not statistically significant in the early HD group (p=0.12) (after adjusting for age and estimated premorbid IQ). In the gene carriers as a whole higher CAG repeat length was associated with greater decline in cancellation time (-0.9 seconds per CAG repeat, 95% CI -1.8, -0.1 seconds, p=0.037) and Ekman fear recognition (-0.2 points per CAG repeat length, 95% CI -0.4, -0.01 points, $p<0.05$), after adjusting for group. A *post hoc* analysis looking at the early HD group alone found that the relationship between CAG repeat length and decline in cancellation time was still statistically significant (-1.1 seconds per CAG repeat, 95% CI -2.3, -0.04, p=0.043) and there was no evidence that the slope of the relationship differed from that found in the gene carriers as a whole. The association with fear recognition was not statistically significant in the HD group alone ($p>0.05$).

There was also a tendency for higher CAG repeat length to be associated with greater decline at category fluency (-0.7 words per CAG repeat length, 95% CI -1.5, 0.05, p=0.066). Higher CAG repeat length was associated with less decline at Ekman disgust recognition (0.2 points per CAG repeat length, 95% CI 0.05, 0.4, $p<0.05$). There were no other statistically significant associations between change in cognitive score and CAG repeat length.

9.3.2.4 Relationship with baseline motor score, disease duration and interaction between CAG repeat length and duration

In general there was little evidence of any relationship between baseline motor score and cognitive change, with most of the associations small and very close to zero, and no statistically significant associations.

There was also little evidence of any relationship between disease duration or age at onset and cognitive change, with the exception of cancellation for which a one-year increase in disease duration (having adjusted for age at baseline) was associated with a one second larger decline in cancellation time (95% CI 0.1, 1.7 seconds, p=0.004). No other associations approached statistical significance.

Finally, there was also little evidence that the effects of CAG repeat length or disease duration on cognition interacted. In general the interaction terms were close to zero and not statistically significant.

9.3.2.5 Relationship with probability of onset

In general there was little evidence of associations between change and probability of onset in the 19 premanifest subjects, with the exception of digit span forwards, for which there was a trend for proximity to onset to be associated with greater decline (1.3 point decrease in change for a 0.1 increase in probability of onset, 95% CI -2.7, 1.1 points, p=0.068). There were no other statistically significant associations between change in cognitive score and probability of onset.

9.3.2.6 Comparison of NART and Spot the Word test

Mean (SD) NART and Spot the Word scores (transformed to be on the same scale as the NART) at both timepoints, with change over time, are shown in Table 9-5. In the cohort as a whole mean NART score increased over time (p=0.0026), whilst mean Spot the Word score did not (p=0.57). There was no evidence that the amount of change in either measure differed between groups.

Table 9-5 Mean (SD) NART and Spot the Word scores at baseline and 12 months for the whole cohort (N=77) with change over time and 95% confidence intervals for the change

	NART	Spot the Word
Baseline	105.4 (11.7)	102.5 (14.0)
12 months	106.4 (11.9)	101.9 (15.9)
Change	1.1 (3.0)	-0.6 (10.0)
95% CI	0.4, 1.8	-2.9, 1.6

Positive change indicates an improvement in score from baseline to 12-month assessment.

The standard deviation of change in Spot the Word score was larger than that for the change in NART score (p<0.0001). However the difference between the amount of change for each measure was not statistically significant (mean difference 1.7, 95% CI -0.5, 4.0, p=0.13).

9.4 DISCUSSION

9.4.1 Cognitive decline in early HD

In this large group of early and premanifest HD subjects followed over one year, cognitive change was variable, and decline in most tasks was very slight. Whilst there was a tendency for the early HD group to decline in all domains except

auditory verbal memory and facial emotion recognition, decline was only statistically significant for the RMF, RMW, TMT B, phonemic fluency and the UHDRS cognitive score (which was possibly driven by the contribution of phonemic fluency). At baseline the early HD group showed impairment across domains (see chapter 4), whilst the premanifest group did not, which suggests that performance in many tasks in the early HD group represents a decline relative to previous levels of ability. Subjects were not scoring at floor level on any tasks, and where possible alternative test versions were used at each assessment in order to minimise practice effects. Hence, although many of the tasks used here were insensitive to the effects of the disease over one year, change in these domains might be detectable over a longer time period.

On the whole the findings in the early HD group confirm those of other studies, a number of which have found decline over time in TMT B and phonemic fluency (Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001; Ho *et al.* 2002; Ho *et al.* 2003; Lemiere *et al.* 2004; Ward *et al.* 2006). Most of these studies also found significant decline in other executive function and psychomotor tasks including Stroop colour and word reading, category fluency, SDMT and TMT A, although with the exception of Snowden *et al.* (2001) the other studies had longer follow-up periods than the current one, which was likely to have increased their ability to detect relatively small effects. In the current study, slight (but not statistically significant) decline was seen on all these tasks, with the exception of category fluency. As mentioned in chapter 4, category fluency was thought to decline at a point at which many subjects were performing at floor on phonemic fluency (i.e. at a later point in the disease process), so this may explain the relative insensitivity of category fluency in this particular cohort. Interestingly, decline

in Stroop colour and word reading was greater in magnitude than decline in Stroop interference, again supporting the theory that a deficit in relatively automatic, well-learned skills (such as word reading) is one aspect of the impairment seen in HD. However, it was in phonemic fluency and TMT B, both of which have a switching component, that statistically significant decline was detected. UHDRS motor score also declined significantly in this cohort, and both TMT B and phonemic fluency have motor components, as well as relying on cognitive switching and (in the case of TMT B) oculomotor skills. It is likely that the combination of skills required for these tasks (visual scanning, motor, switching, working memory) and their known reliance on striatal and frontal regions makes them particularly sensitive to decline in this cohort, even over the relatively short time period used here.

There have been mixed findings in relation to digit span decline in early HD, with some studies reporting a decline (Snowden *et al.* 2001; Ho *et al.* 2003) and others no change (Bachoud-Lévi *et al.* 2001; Lemiere *et al.* 2004). In this study change in early HD subjects was not significantly different to that seen in controls. Subjects did decline, on average, very slightly at forwards and backwards digit span. Notably they did not display the slight practice effects seen in controls, suggesting that digit span might have the potential to detect change with larger numbers or over a longer time period.

Studies that have used the HVLT (or the similar Rey Auditory Verbal Learning Test) have tended to find no change over periods of up to four years in early HD (Bachoud-Lévi *et al.* 2001; Ho *et al.* 2003; Lemiere *et al.* 2004), an exception being Ward *et al.* (2006) who found that HVLT immediate recall declined over

four years. Similarly in this study there was no evidence of decline in the early HD group relative to controls on any subtest. However some subjects in all groups scored at ceiling on delayed recall and discrimination at both timepoints, meaning that these two subtests probably lacked sensitivity. In contrast, the early HD group showed, on average, significant decline at both subtests of the RMT. Although on the RMW some of the controls were at ceiling at both timepoints, meaning that the variance of change in controls was likely to have been reduced and perhaps lead to over- or under-estimation of the effect of group differences, none of the early HD subjects scored at ceiling at follow-up; also no subjects were scoring at ceiling in the relatively harder RMF. As discussed in chapter 4, the RMT differs from the HVLT in that recognition for 50 items is tested with a two-alternative forced choice paradigm, rather than a yes / no decision to separate target and distractor stimuli. Baseline data from the HVLT suggested that whilst recognition of true positives might be reduced, but reasonable given the number of words encoded, early HD subjects still tended to make more false positives than controls. The recognition task in the RMT is relatively harder, as more stimuli must be recognised correctly, and the subject must choose between two stimuli (one true positive, one false positive) both of which he may think he has seen before. This task may therefore be more susceptible to increasing fronto-striatal dysfunction.

In contrast to the RMW, group differences in change on the RMF were due mainly to improvement in controls and very little real decline in the early HD group (although as some of the controls were at ceiling at follow-up on the RMW there may also have been a practice effect on the latter task which was not detected). Whilst early HD subjects were not at floor level on the RMF at either

timepoint, they were already performing relatively worse at the RMF than the RMW at baseline (apparent from both age-scaled scores and the effect sizes of the group differences at baseline) suggesting that decline at this task might occur earlier (and perhaps faster) relative to decline at the RMW. The fact that baseline differences are concordant with the change scores suggests that the tests are reflecting real disease-related deficits.

As discussed in chapter 5, facial recognition ability is affected by eye movements, which are known to be abnormal even early in the disease, so this might partly explain a different time-course of decline in the two skills. Also, since there was no evidence of impairment in the far-from-onset premanifest group, but impairment is clear in the early HD group, there may be a time (perhaps in closer-to-onset premanifest subjects, or around the time of phenoconversion) when this test is sensitive to change in this population.

In general the memory deficit in HD would benefit from further probing; whilst performance on auditory verbal learning tests tends not to show decline, story and object recall (Snowden *et al.* 2001), visual span (Bachoud-Lévi *et al.* 2001), spatial span (Ho *et al.* 2003) and digit span forwards and backwards (Snowden *et al.* 2001; Ho *et al.* 2003) have all shown decline in some early HD populations. Given that some of the early HD subjects in this study were at ceiling on some subtests of the HVLT, it may be that test difficulty needs to be manipulated by increasing the number of items to be remembered, or the number of distractor items in recognition paradigms, in order to detect small changes in ability.

Performance on the GNT and VOSP was not impaired at baseline and did not show significant decline over one year, although as discussed in chapter 4, these

tasks tap single aspects of naming and visual perception and hence these findings do not preclude subtle decline in language or visuoperceptual abilities. In general others have also failed to find evidence of significant decline in these skills in early HD (Bachoud-Lévi *et al.* 2001; Ho *et al.* 2003) although one study, with follow-up of just under three years, found that both naming and VOSP performance declined (Lemiere *et al.* 2004), perhaps indicative of the fact that decline is relatively slow in these domains.

Performance on the Benton and the Ekman facial emotion recognition battery was impaired at baseline, but there was no statistically significant change in these skills over one year, again perhaps reflecting a slow rate of change in these skills, or a lack of test sensitivity to change. The version of the Ekman used here had only four exemplars of each emotion, which could have contributed to practice effects and made it harder to detect decline. Only one study to date has investigated decline in emotion recognition, and this was in premanifest, rather than early HD subjects, in which there was no evidence of decline over one year (Sprengelmeyer *et al.* 2005). Given the potential impact of changes in emotion recognition skills on patients' and carers' lives more work needs to be done to clarify whether, and how, these deficits change over time.

9.4.2 Cognitive decline in premanifest HD

There was very little difference between change in the premanifest group and that seen in controls over the time period investigated here. As discussed in previous chapters, the premanifest subjects presented here were on average 18 years from predicted motor onset (with most over 10 years from onset) and the rate of decline in many domains is thought to be negligible until within the last

decade prior to onset (Paulsen *et al.* 2007). On many tasks this group showed slight improvements over time. Practice effects tend to be largest between the initial test and the next follow-up (Bachoud-Lévi *et al.* 2001) which might have played a role in this although alternative test versions were used where they were available. Practice effects seemed similar in the premanifest groups and controls, reinforcing the fact that in this cohort there was little to distinguish the two groups, at least using current clinical or cognitive scales. With the exception of studies in which premanifest subjects have phenoconverted during the assessment period (Paulsen *et al.* 2001; Snowden *et al.* 2002), change in premanifest subjects has rarely been detected (Lemiere *et al.* 2002; Lemiere *et al.* 2004; Witjes-Ane *et al.* 2007).

In the current study phonemic fluency was the only task on which the premanifest group showed significant decline relative to controls. In fact both groups showed slight improvement (perhaps due to practice effects), but the improvement was far greater in the control group. One other study has reported a decline in phonemic fluency performance in 70 premanifest subjects who had converted over the two years of the study, relative to those who had not (Paulsen *et al.* 2001). In that study the converters also showed significant decline at the SDMT and the three Stroop subtests. Others have failed to detect decline, although some have summed the words generated for three different letters for a minute each (as was done here), whilst others have used just one letter for one minute, and so test differences may have contributed to the differences between studies, (compare e.g., Snowden *et al.* 2002; Witjes-Ane *et al.* 2007). One slightly smaller study found a similar pattern of results to those seen here, over

an average of 2.5 years, but group differences were not statistically significant (Lemiere *et al.* 2004).

As discussed in chapter 5, phonemic fluency has a switching component, which is thought to rely more on frontal parts of fronto-striatal circuits, but also has a motor component, and phonemic switching ability has been shown to decrease with increasing motor impairment in HD (Ho *et al.* 2002). On average, UHDRS motor score increased slightly in the premanifest subjects over the course of the study; this change approached statistical significance and may have had clinically relevant effects on function. Phonemic fluency was the only executive function task for which premanifest performance at baseline approached early HD levels, so together these results suggest that the combination of demands of this task might be such that decline in performance can be detected very early in the disease course. Given that not all studies have found decline in phonemic fluency in premanifest subjects, further longitudinal work is needed, to clarify whether decline is detectable a certain number of years prior to motor onset, or in subjects with particular levels of motor ability.

It is notable that on some tasks the early HD group declined more relative to controls, whilst the change in the premanifest group was not statistically significantly different to that in either of the other two groups, suggesting that their performance was lying somewhere between the two. Other work in this cohort has shown that both caudate (see chapter 5) and whole-brain volumes (see chapter 4) are reduced in this relatively far-from-onset group, and hence it is not unreasonable to assume that some cognitive sequelae may be detectable given sensitive enough tests.

9.4.3 Associations between change and clinical variables

In the premanifest group there was little evidence that proximity to onset was related to a faster rate of decline in cognitive tasks. Even for digit span forwards, for which there was a trend towards a significant association, performance was generally stable or improved over time, with only five premanifest subjects declining. It seems unlikely that these findings would be replicable, particularly since other evidence suggests that memory performance declines around the time of onset (Snowden *et al.* 2002) and that in general decline in premanifest subjects does not become obvious until within 10 years of predicted onset (Paulsen *et al.* 2007), whereas the current cohort was on average 18 years from estimated onset.

CAG repeat length was not associated with change in most variables, and there was also little evidence that either baseline motor score or disease duration was predictive of rate of cognitive decline in any domains. Although higher CAG repeat length was associated with faster decline at cancellation, facial fear recognition, and digit span forwards (the latter in the premanifest group only), the lack of association in other similar tasks makes this pattern hard to interpret. Higher CAG repeat length was not associated with longer disease duration in the early HD group so the associations cannot be explained simply as an effect of the latter. In general the amount of cognitive decline was so slight and variable, that it is unsurprising that change was not associated with these demographic variables. As mentioned in section 1.4.1, there are likely to be many other factors contributing to cognitive performance and change over time, which may mask the (probably small) effects of the variables investigated here. Others have also found little effect of repeat length on rate of cognitive decline (Bachoud-Lévi *et al.* 2001; Snowden *et al.* 2001; Ward *et al.* 2006). However the two

larger studies, one of which followed subjects for four years, have found that disease duration or indices of motor function explained some of the variance in change scores (Snowden *et al.* 2001; Ward *et al.* 2006), with the suggestion that motor score might better reflect disease severity than duration. Others have demonstrated small but statistically significant effects of both CAG repeat length, disease duration, and their interaction on decline across domains (Rosenblatt *et al.* 2006) although cognition was measured using the Mini Mental State Examination (MMSE) (Folstein *et al.* 1975) which does not provide a very specific measure of many of the skills known to be impaired in HD. Ideally the effects of these variables on change in individual tasks still needs to be investigated in large cohorts over longer time periods than that used in this study.

9.4.4 Comparison of NART and Spot the Word Test

Finally, following the suggestion that the Spot the Word Test might be a more suitable estimate of premorbid IQ than the NART in HD, because of the dysarthria associated with increasing progression of the disease, the ability of the two tests to produce stable estimates of premorbid IQ over time was investigated. Although the NART showed significant change (increase) over time whilst the Spot the Word test did not, there was no evidence that the amount of change differed between the tests. In fact the 95% confidence intervals for change in Spot the Word demonstrate that it could well be of a similar magnitude to that seen in the NART, but the greater variance associated with change in Spot the Word meant that the change was not statistically significant. Overall, the lack of difference in the amount of change between the two measures, and the reduced variance associated with change in the NART, suggests that the NART gave a more stable estimate of premorbid IQ in the cohort presented here.

9.5 CONCLUSION

This work confirmed that of a number of others, demonstrating that cognitive change in premanifest and early HD was slight and variable, particularly over the one-year time period used here. However, even over this short time period the early HD group showed decline in two executive function tasks and a memory task, and there may be potential for more sensitive tasks to be developed, particularly in the domain of memory. In the premanifest group change in phonemic fluency was also reduced relative to controls, and as cognitive change is so hard to detect in this widely variable population this finding merits further investigation, in order to see whether it is replicable and if so, how this depends on the characteristics of the premanifest population under investigation.

Although there is some evidence that rate of decline is determined at least in part by CAG repeat length and disease duration, there was little evidence to support that in the current study. However, it seems likely that such effects exist, but that they are relatively small, making associations hard to detect over short time periods and particularly in the domain of cognition.

One of the aims of this work was to investigate potential markers of disease progression. Problems with cognition as a marker were discussed in the introduction, predominantly its susceptibility to practice effects and other non-disease-related factors, all of which contribute to its variability. However, unlike markers than can be measured more objectively, such as brain volume or "wet" biofluid markers, cognitive tests provide a direct measure of the day-to-day function of patients. It is these outward signs of the disease that patients and carers notice and find difficult; further understanding of these changes will help

with disease management, and potential therapies for the symptomatic population will need to have a positive effect on function (i.e., at the very least prevent further decline) in order to improve the quality of life of people with HD. Hence, despite the fact that the change seen here was, as in other studies, small and variable, it is important that further work continues to focus on aspects of cognitive change in HD, and how it can best be measured.

10 MEASURING WHOLE-BRAIN ATROPHY IN HUNTINGTON'S DISEASE: AGREEMENT BETWEEN THE BBSI AND "TRUE" VOLUME LOSS

10.1 INTRODUCTION

The BBSI measures the shift in intensity within a specific window, chosen to encompass the boundary under investigation (the CSF/brain boundary, which is mainly CSF/grey matter) (see section 2.5.4.3). The intensity window is described in terms of its centre and width, and these parameters can be adjusted depending on the typical intensities of the set of images being investigated.

The changes in intensity across the CSF/brain boundary, and the intensity window within which the BBSI is calculated, are demonstrated in Figure 2-5, page 85. Chapter 11 reports whole-brain atrophy rates in HD, measured using the BBSI, as well as the associations between atrophy rate and the cognitive and clinical changes described in the previous chapter. This chapter describes how the BBSI parameters were adjusted to get the best agreement between volume loss estimated by this technique, and estimates of the "true" volume loss. The final parameters were then used to generate the BBSIs from which the atrophy rates in the following chapter were calculated.

In T1-weighted images CSF appears dark (low intensity), grey matter lighter (medium intensity) and white matter bright (high intensity). Scans are normalised to mean white matter intensity prior to BBSI calculation meaning that intensities vary from 0 to 1, with CSF typically around 0.2 and GM around 0.8. The default window for BBSI calculations has a centre of 0.5 with a width of 0.5, i.e. it is designed to fall in the middle of the CSF and grey matter intensities and

cover a range of intensities around this. This reflects the original validation which was based on a particular acquisition and scanner.

The optimal parameters are those for which the BBSI most closely agrees with the "real" volume difference and these will vary with different scan intensities. If the BBSI measured the volume change for two scans exactly, then the mean and standard deviation of the difference between the BBSI and the real change would be zero. In practice this is rarely the case; the BBSI has been shown to underestimate change slightly (Freeborough and Fox 1997) so the mean difference is unlikely to be zero. This is because even the optimal window centre and width will not encompass the boundary perfectly on all scans, due to variability between scans, and also because in some areas there is no "boundary" in terms of signal intensity, e.g., brain stem or areas where cortex overlaps or is contiguous with dura or other non-brain tissue of a similar signal intensity. Also, the "gold standard" measure of volume change is the difference between two manual segmentations and both manual segmentation and the BBSI are subject to error, so the difference between the two is unlikely to have a standard deviation of zero. The agreement between the BBSI and manual measures is assessed by looking at the standard deviation (SD) of the difference between the two measures (Bland and Altman 1986). The mean of this difference is a measure of the bias of the BBSI compared to manual measures (Bland and Altman 1986). Therefore in order to maximise agreement between the BBSI and manual measures, the BBSI is calculated for a set of registered scan pairs using different window centres and widths, to find the window centre and width which minimises the standard deviation of the difference between the two measures.

This chapter describes three methods of working out the "best" parameters for the BBSI: the first method derives parameters directly from the scan intensities; the second and third look for the best agreement between the BBSI and two sets of manual measurements. BBSIs from each method are then compared with manual measurements to see which agrees best. Finally, when the final BBSI parameters are chosen, the agreement between these measurements of loss and manual measurements is discussed.

10.2 METHOD 1: DETERMINING PARAMETERS FROM FIRST PRINCIPLES

The optimum window centre should lie approximately half-way between the mean CSF intensity and mean grey matter intensity, since it is change across this window that will represent a boundary shift. Tissue intensities are normalised to the mean white matter intensity prior to calculation of the BBSI, so samples of white matter, grey matter and CSF were taken from three pairs of images (baseline and follow-up) in order to calculate representative normalised intensities. The subjects were chosen to cover a range of age, gender and status: a young female control, a middle-aged male premanifest subject and an elderly female HD subject. For white matter and CSF, 64-voxel cubes (4 voxels deep in each plane) were drawn manually, in similar regions for each image: the right frontal white matter and the left lateral ventricle respectively. Cubes were inspected visually to ensure that no other tissue type intruded. Because of the convoluted nature of the cortical grey matter perfect cubes were not sufficient to sample grey matter without intrusion from other tissue types. Two-dimensional regions were drawn manually on four consecutive coronal slices of the left insula to give a three-dimensional grey matter region of 76 voxels. The mean

intensities for each tissue type and each scan were measured from these regions and were normalised by dividing by the mean white matter intensity.

10.3 RESULTS 1

The mean (SD) normalised intensities were CSF 0.20 (0.02) and grey matter 0.65 (0.04) giving a hypothetical window centre of 0.425 (half way between the mean intensities of the two tissue types) (Table 10-1). Based on this a set of window centres around 0.425 (from 0.3 to 0.5 in steps of 0.05) was used to generate the BBSIs reported in the following sections.

Table 10-1 Raw and normalised CSF and GM intensities, with overall mean and range, for the images used for parameter estimation

Subject	CSF (mean, range)	GM (mean, range)	WM (mean)	CSF/WM (mean, range)	GM/WM (mean, range)
1 baseline	26 (20 – 36)	84 (73 – 96)	120	0.22 (0.17 – 0.30)	0.70 (0.61 – 0.80)
2 baseline	26 (18 – 33)	80 (67 – 93)	124	0.21 (0.14 – 0.27)	0.64 (0.54 – 0.75)
3 baseline	24 (14 – 35)	77 (66 – 88)	125	0.19 (0.11 – 0.28)	0.61 (0.53 – 0.70)
1 12 months	24 (16 – 32)	84 (72 – 100)	119	0.20 (0.13 – 0.27)	0.71 (0.60 – 0.84)
2 12 months	28 (20 – 33)	81 (63 – 93)	125	0.22 (0.16 – 0.26)	0.64 (0.50 – 0.74)
3 12 months	22 (14 – 34)	78 (65 – 90)	127	0.18 (0.11 – 0.27)	0.61 (0.51 – 0.71)
			Mean	0.20 (0.11 – 0.30)	0.65 (0.50 – 0.84)

10.4 METHOD 2: DETERMINING PARAMETERS FOR A SUBSET OF IMAGES

A subset of images was chosen to be representative of the set as a whole (see chapter 11 for details of the 62 scan pairs). Five subjects from each of the three groups (controls, premanifest and early HD) were chosen, and paired t-tests or Fisher's exact test were used to ensure that the mean age, gender, CAG repeat length and amount of atrophy (measured by the default BBSI) for the five subjects chosen were not significantly different to those for the subjects not included in each group. In addition these variables were not significantly different for the entire subset of 15 compared with the remaining 47 subjects. Demographic data for this subset and remaining cohort are shown in Table 10-2.

Table 10-2 Demographic data for the optimisation subset and remaining cohort

	Optimisation subset (N=15)	Remaining cohort (N=47)
Gender (M:F)	7:8	23:24
Age (year)	44.0 (12.4)	45.4 (9.1)
CAG repeat length[a]	43.2 (3.0)	43.7 (2.5)
Atrophy (ml)[b]	3.9 (6.1)	5.0 (6.6)

Data are shown as mean (SD) with the exception of gender.

a For 10 gene carriers in the optimisation subset and 34 in the remaining cohort

b BBSI with an interval of 12 months, estimated using the default BSI parameters, centre 0.5, width 0.5

The baseline and registered-repeat images for each of these 15 subjects were segmented manually (according to the protocol in Appendix 2), and the volume difference was calculated; this volume difference should not be influenced by

scanner drift affecting voxel size, since registration (which incorporates scaling factors) should correct for such changes (Whitwell *et al.* 2004). The BBSI was calculated using a range of window centres and widths and for each window centre and width combination the standard deviation of the difference between the BBSI and manually-estimated volume difference was calculated in order to see when it reached a minimum.

10.5 RESULTS 2

There was no evidence of statistically significant differences in demographic variables between the optimisation subset and the remainder of the cohort, or between the five subjects selected from each group and the remainder of each group.

Figure 10-1 shows the standard deviations of the difference between the BBSI and volume difference, for a range of window widths and centres.

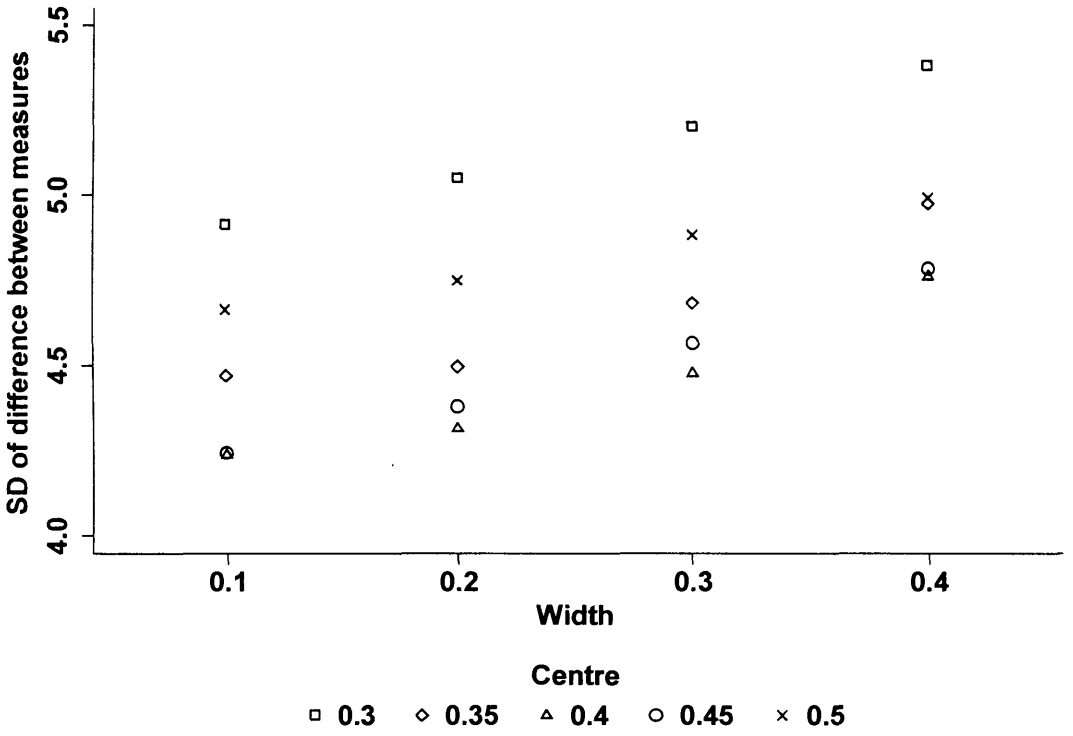


Figure 10-1 SDs of differences between volume change estimated by the BBSI and manual measures for a range of window widths and centres (subset)

The centre that minimised the SD (and therefore was showing best agreement between measures) was between 0.4 and 0.45, and so further analyses were run with centres varying from 0.4 to 0.44 in steps of 0.01. The width was kept constant at 0.2, which was preferred to 0.1 since a width of 0.1 might cause the boundary to fall outside the window due to variations in scan quality, resulting in a more variable BBSI measurement.

Figure 10-2 shows the SDs of the difference between measures for these parameters, the best of which was 0.41. At this stage a change of window centre by 0.01 was changing the SD very little and so no other parameters were tested.

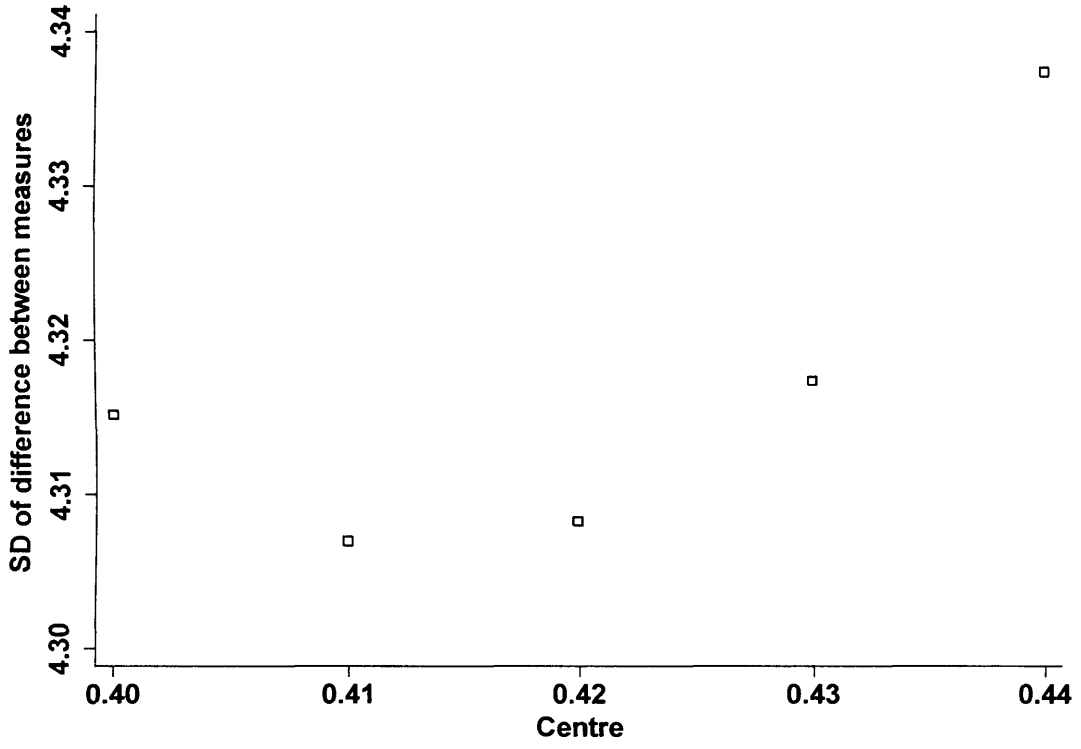


Figure 10-2 SDs of differences between volume change estimated by the BBSI and manual measures for a small range of centres (width 0.2) (subset)

10.6 METHOD 3: DETERMINING PARAMETERS FOR ALL IMAGES

All 124 images (62 baseline and 62 follow-up) were segmented in native space prior to registration. The volumes of these segmentations were used to calculate change in volume between baseline and follow-up. The registered scan pairs had the BBSI recalculated a series of times using the same range of window centres and widths as in section 10.4. Agreement between measures was again assessed by looking at the standard deviation of the differences between them. Whilst this method has the advantage that it includes the whole dataset, the volume

difference against which the BBSI is compared is calculated from non-registered images. As the images are taken on average one year apart it is possible that scanner changes during the intervening months could contribute to the measured difference (Whitwell *et al.* 2004), as well as real change (and segmentor error). Thus whilst this might still yield the same optimal parameters as method 2, the mean differences between the measures will not be a good estimate of the amount by which the BSI and "true" volume change differ.

10.7 RESULTS 3

Figure 10-3 shows the SDs of the difference between measures for a range of window widths and centres.

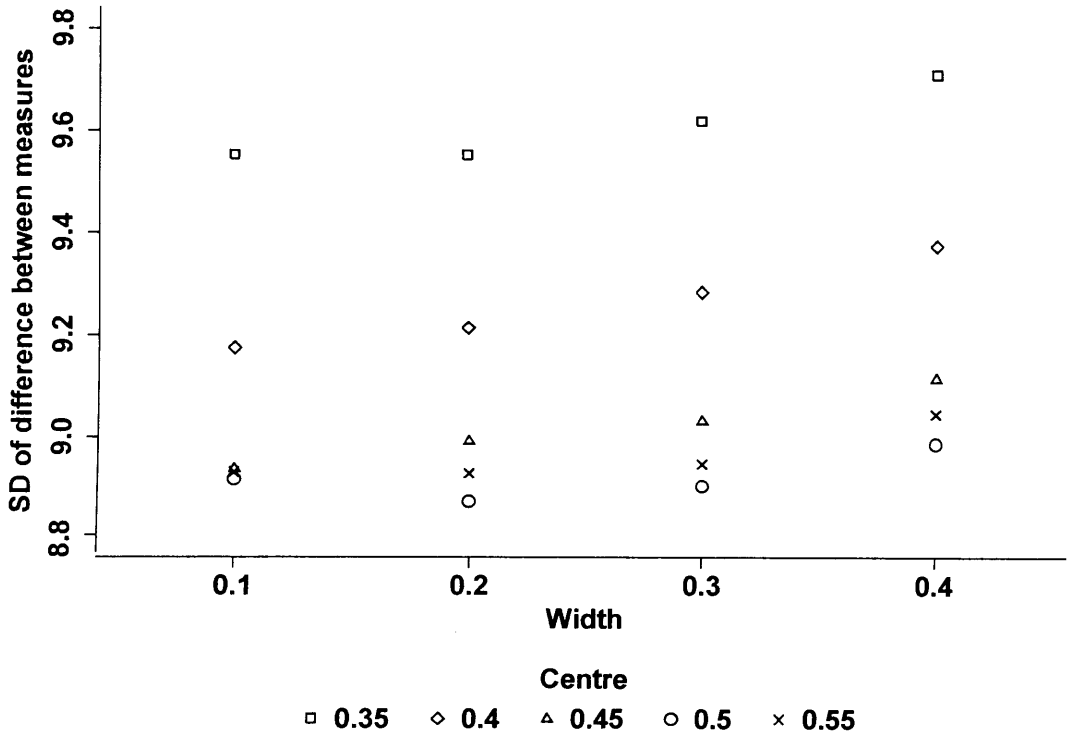


Figure 10-3 SDs of differences between volume change estimated by the BBSI and manual measures for a range of window widths and centres (entire cohort)

The centre that minimised the SD was between 0.5 and 0.55, with a width of 0.2 and therefore further analyses were run with centres varying from 0.5 in steps of 0.01, until the SD started to increase (keeping the width constant at 0.2). Figure 10-4 shows the SDs of the differences between measures for these different centres, the best of which was 0.51. At this stage a change of window centre by 0.01 was again changing the slope very little and so no other parameters were tested.

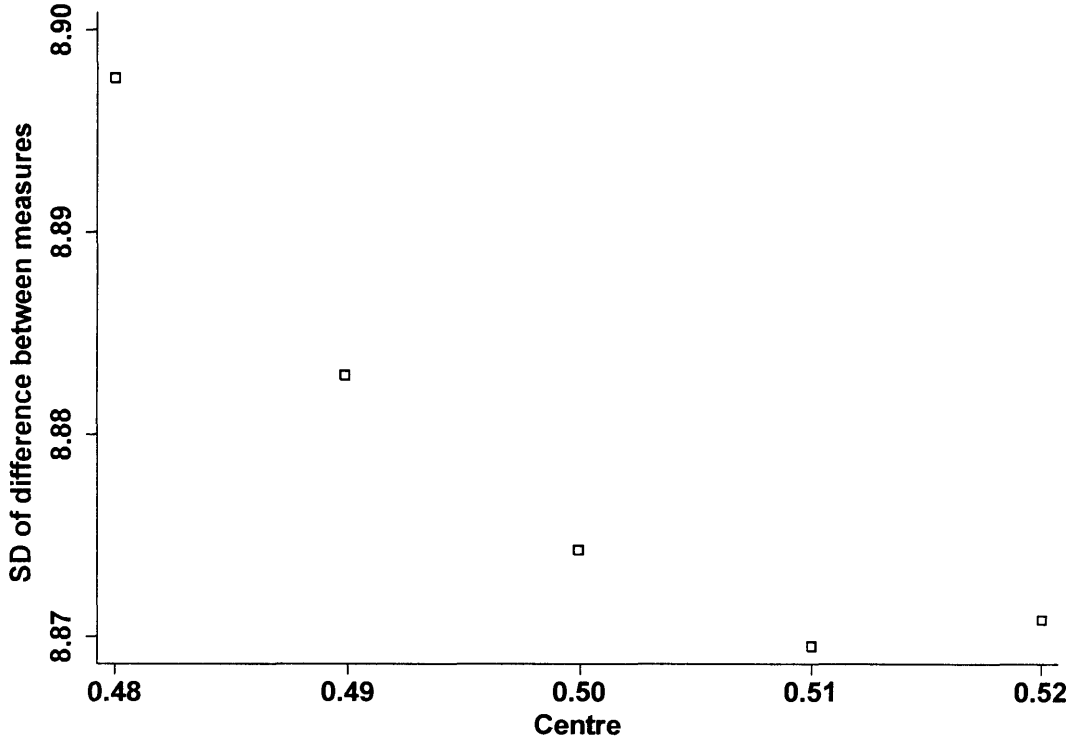


Figure 10-4 SDs of differences between volume change estimated by the BBSI and manual measures for a small range of centres (width 0.2) (entire cohort)

10.8 COMPARISON OF THE THREE SETS OF OPTIMISATION PARAMETERS

Thus far the parameters have been chosen either because they maximised agreement between the BBSI and manual measures (methods 2 and 3) or because

they were derived from looking directly at scan intensities (method 1). Two sets of manual measures have been used, using segmentations of registered scan pairs (on the subset) and segmentations from non-registered scan pairs (on the whole cohort) with the caveat that in the latter case the manual measures are likely to be subject to more error because of scanner change over time.

In order to compare the three sets of parameters directly each set was used to generate BBSIs for the subset, so that it could be compared against the manual measures which are the best estimate of volume change available. The mean difference between each set of BBSIs and manual measures, and the SD of this difference, are show in Table 10-3.

Table 10-3 Parameters from the three methods, with the mean and SD of the difference between each set of BBSIs and manual measures (subset)

Optimised on:	Intensities	Subset	Entire cohort
Window centre	0.425	0.41	0.51
Window width	(0.2)	0.2	0.2
Mean of difference (ml)	-3.23	-3.30	-3.40
95% CI	-5.62, -0.84	-5.68, -0.91	-6.08, -0.71
SD of difference	4.32	4.31	4.85

As discussed in section 10.1, the ideal mean difference between the BBSI and segmented volume difference is zero, but in practice the BBSI is known to underestimate volume loss slightly (Freeborough and Fox 1997). All three sets of parameters were significantly underestimating the true volume difference, and the parameters which led to the least bias were those derived from image intensities (-3.23 ml). The parameters which led to the best agreement were

those derived from the subset of images (SD of 4.31). Pitman's test was used to see whether the standard deviations of the difference differed between these two sets of parameters, but there was no evidence to support this (ratio of standard deviations=0.997, 95% CI 0.97, 1.02, p=0.79). A paired t-test was used to see whether the bias (mean difference) was significantly different between the two sets of parameters but again there was no evidence to support this (mean difference=0.07, 95% CI -0.04, 0.17, p=0.18).

In the light of these findings it was decided to proceed with the window centre derived from the image intensities (0.425); BBSIs showed the least bias, although marginally, and the agreement between BBSIs and manual measures was not significantly different to the centre of 0.41 (which showed the best agreement). A window width of 0.2 was chosen rather than 0.1 because although the narrower window tends to increase agreement there is also more chance that a poor scan will fall completely outside the window and thus fail to have the boundary shift estimated accurately. The BBSIs reported in the rest of this chapter (and in chapter 11) are thus derived from the final parameters of centre 0.425, width 0.2.

10.9 COMPARISON OF THE BBSI AND MANUAL MEASURES

Volume loss as estimated by segmentation and the BBSI, for the entire cohort, is shown in Figure 10-5. As volume loss from segmentation is estimated from the images in native space, and so is subject to scanner drift in addition to segmentor error, the mean differences between each measure look greater than the 3.23ml it was shown to be (see section 10.6 for discussion of using non-registered images

to estimate volume loss and section 10.8 for derivation of the mean difference between volume loss and the BBSI).

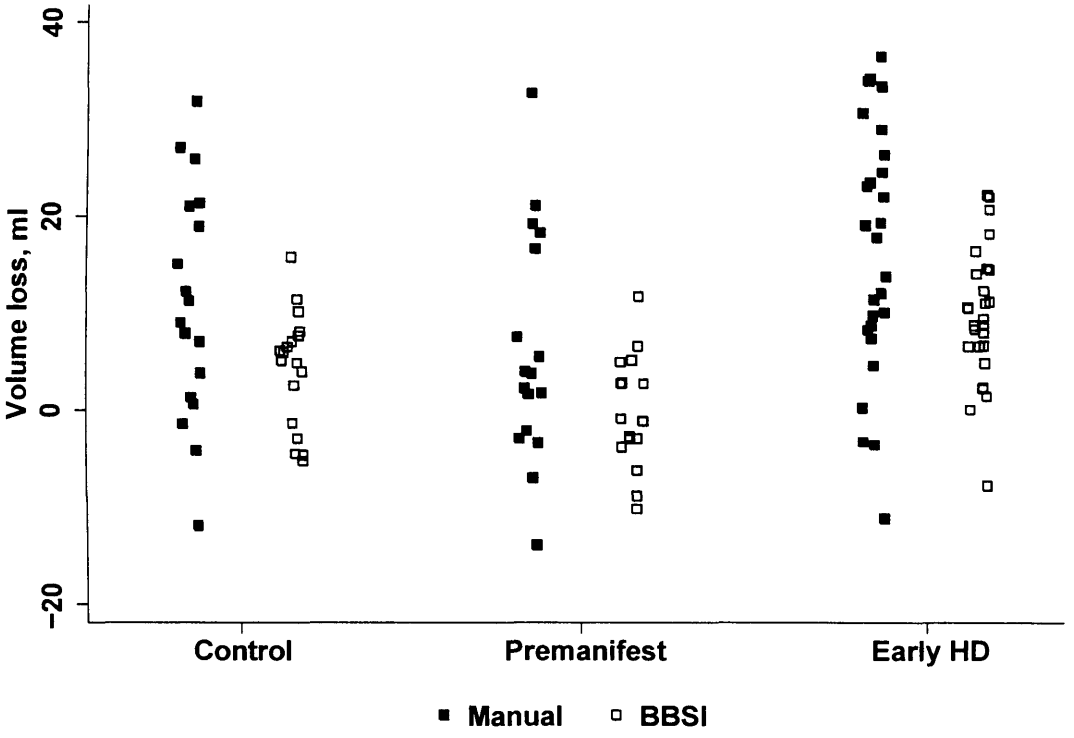


Figure 10-5 Volume loss for each group measured manually and with the BBSI

Pitman's test was used to compare the variance of each measure within each group, and in all cases use of the BBSI reduced group variance compared to the manual measure (controls: ratio=2.09, 95% CI 1.52, 2.86, $p<0.0001$; premanifest: ratio=2.05, 95% CI 1.43, 2.94, $p<0.0001$; early HD: ratio=1.86, 95% CI 1.34, 2.59, $p<0.0001$).

Linear regression models were used to investigate differences between groups, for each measure. Differences between groups were slightly larger for volume change estimated by the BBSI (see Table 10-4).

Table 10-4 Mean (SD) volume loss from manual and BBSI measures, with group differences

	Control (N=18)	Premanifest (N=17)	Early HD (N=27)	Difference (Premanifest - Control)	Difference (Early HD - Control)
Manual (ml)	11.0 (11.9)	6.2 (11.9)	16.2 (12.7)	-4.8 (p=0.25)	5.2 (p=0.17)
95% CI	5.1, 17.0	0.05, 12.3	11.2, 21.2	-13.1, 3.4	-2.3, 12.6
BBSI (ml)	4.5 (5.7)	-0.5 (5.8)	9.8 (6.8)	-5.0 (p=0.022)	5.3 (p=0.007)
95% CI	1.6, 7.3	-3.5, 2.5	7.1, 12.4	-9.2, 0.7	1.5, 9.1

Positive differences indicate greater volume loss in HD subjects relative to controls.

10.10 DISCUSSION

In this section I have discussed three methods of finding the BBSI parameters which give best agreement between this measure and manual measures, which are the current "gold standard".

Two methods gave similar parameters: deriving parameters from intensities, and deriving parameters from a subset of scans where the manual measure was calculated from registered pairs. The first method, based on image intensities, was only based on six images suggesting that even this small sample size was sufficient to obtain a reasonable estimate of the window centre and width. This requires subject selection, and careful manual outlining of the regions of interest. It is likely to be influenced by the placing of the regions of interest, particularly in images with some inhomogeneity. The second method compared the BBSI with the best "gold standard" measurement, manual measures from registered scan pairs. This method is time-consuming, again requiring careful subject selection in addition to resegmentation of the registered images, but having done this it allows the best estimate of agreement and bias.

The third method used the whole cohort which is the most practical method, since all the images have been segmented in native space prior to registration, but parameters derived from this were quite different to those from the other two methods and were associated with greater bias and less agreement than the previous two methods, almost certainly because of the added error associated with manual measures derived from non-registered scan pairs.

The fact that the first two methods give very similar results is encouraging, and it suggests that for future work it may be sufficient to use either method to get a set

of parameters which will maximise agreement. The third method was not very satisfactory. In a large-scale study which method is chosen will depend on practicalities such as time constraints and segmentor availability.

All agreement improved as the window widths narrowed, which suggests that the set of scans here were uniform enough that the intensity changes associated with a shift of border were all contained within quite a narrow window. This is possibly due to the fact that all the images here were from the same scanner and a single site, with a relatively short and invariant time interval. A window width of 0.1 seemed to work for most centres; however it did not produce the best agreement for some centres, suggesting that at this width some noisy images might be falling outside the range. Hence a window width of 0.2 was chosen for the final parameters.

10.11 CONCLUSION

This chapter discussed three ways of calculating BBSI parameters which maximised agreement between this measure and the best estimate of true volume change (manual segmentation). Parameters estimated from scan intensities, or from volumes derived from segmented registered scan pairs were similar, whilst those from non-registered scan pairs were different and led to the BBSI agreeing less well with manual measurements. The bias for the parameters finally chosen was -3.23ml (95% CI -5.62, -0.84) suggesting that the BBSI was underestimating volume loss slightly. However, as has been found previously, the BBSI significantly reduced within-group variance and also slightly increased group separation, compared with manual measures (Freeborough and Fox 1997).

BBSIs from these parameters were used to calculate the atrophy rates that are presented in the following chapter.

11 LONGITUDINAL CHANGE IN HUNTINGTON'S DISEASE: APPLICATION OF THE BBSI TO MEASURE CEREBRAL ATROPHY RATE AND ASSESS ITS ASSOCIATIONS

11.1 INTRODUCTION

There have been relatively few longitudinal structural MRI studies in HD, and the majority of them have focused on the striatum. However, cross-sectional work, including that reported in earlier chapters, suggests that extra-striatal atrophy occurs early on in the disease, and that change in volume can be detected in these regions (e.g., Kipps *et al.* 2005). The work in chapter 3 demonstrated that increased rates of whole-brain atrophy could be detected in HD, and that the BBSI was a useful technique for this measure. However the study described there had a small sample size and took place over a short period of time (six months) and thus part of the aim of the following work was to replicate and expand on those findings in a larger cohort, over a longer time scale. In addition to investigating whether the finding of increased rates of whole-brain atrophy could be replicated, this work also focused on whether this measure could be a potential biomarker of progression in HD, by looking at associations with genetic and demographic variables, as well as whether atrophy rate was predictive of decline in other domains.

11.2 METHODS

11.2.1 Subjects and assessments

The 77 subjects who returned for the follow-up assessment were also asked to have a repeat MRI scan (18 controls, 19 premanifest, 40 early HD, see section 9.2). At the follow-up assessment two back-to-back volumetric images, with the

same parameters, were acquired. This increased the chances of obtaining at least one image free of motion or other artefact at the time of the assessment and reduced the risk that subjects would need to be called back for rescans.

One early HD subject was unable to be scanned at follow-up due to anxiety and three early HD subjects had chorea that affected image quality. Thus longitudinal MRI data were initially acquired for 18 controls, 19 premanifest subjects and 36 early HD subjects.

11.2.2 Image segmentation and the BBSI

Whole brains were segmented following the protocol in Appendix 2. Follow-up scans were then positionally matched to their baseline images using an affine registration with 12 degrees of freedom (in which the cost function minimized was the standard deviation of the ratio image, see section 2.5.4.2) (Woods *et al.* 1998). The BBSI was calculated from the registered scan pairs as described in section 2.5.4.3, using the parameters determined in chapter 10.

Registered scan pairs were checked by an experienced imager with no knowledge of the subjects' disease status or atrophy rates. Eleven subjects were deemed to have inadequate scan pairs for analysis because of movement or scanner artefact (two premanifest and nine early HD). Thus 18 controls, 17 premanifest subjects and 27 early HD subjects were included in the longitudinal MRI analysis. Demographic data for the included and excluded subjects were subsequently compared to investigate whether exclusion was related to factors such as disease status or chorea score.

At baseline 53 subjects had all assessments on the same day or within three days, four subjects had rescans after the clinical and cognitive assessments (mean (SD) interval 5.7 (2.6) weeks, range 2 – 7 weeks), and five subjects had the clinical assessment after the other two for logistical reasons (mean (SD) interval 11.5 (6.1) weeks, range 7 – 22 weeks). At follow-up 58 subjects had all assessments on the same day or within two days, three subjects had clinical assessment after the other two for logistic reasons (mean (SD) interval 10.2 (7.2) weeks, range 5 – 18 weeks) and one subject completed the cognitive assessment but was claustrophobic and had to return to complete the scan and clinical assessment nine weeks later.

11.2.3 Statistical analysis

t-tests assuming unequal variance were used to investigate group differences in age, estimated premorbid IQ and CAG repeat length at baseline. A chi^2 test was used to compare gender in each group. Fisher's exact test was used to see if the proportion of right-handed subjects differed between groups.

Atrophy rates were analyzed on a logarithmic scale according to equation 11.1 in order to ensure that doublings and halvings be treated as effects of equal magnitude; follow-up volume was calculated from the BBSI.

$$\left[\frac{\left(\ln\left(V_{follow-up}/V_{baseline}\right)\right)}{time}\right] \qquad 11.1$$

Regression models relating the log-transformed variables to subject group and age were used to obtain individual age-standardized atrophy rates. Geometric mean atrophy rates (as annualized percentage change from baseline, % per year)

were calculated by back transformation with standard deviations (SDs) calculated from variance transformation formulae.

For cognitive and clinical change a simple difference score was calculated for scores at the two timepoints such that a negative difference score signified decline (see chapter 9). The majority of the tests were timed or open-ended (subjects could provide as many answers as possible) and on these there was no evidence of floor or ceiling effects. As discussed in chapter 9, some subjects in all groups were scoring at the maximum at both timepoints for HVLT delayed recall and discrimination, and the RMW, and hence these scores may not be sensitive to change over the time period employed here.

11.2.3.1 Group differences

A linear regression model with robust standard errors was used to compare rates of whole-brain atrophy (% per year) between groups. Although the effect of age on atrophy rates is thought to be relatively small in this age range (Scahill *et al.* 2003), age at baseline was included as a covariate in this model in order to ensure that any group differences seen could not be attributed to this.

11.2.3.2 Associations with baseline clinical measures

Separate linear regression models were used to assess the association between atrophy rate and probability of onset (in the premanifest group) and disease duration, UHDRS independence score and TFC (in the early HD group), adjusting for age at baseline.

As in previous chapters a linear regression model (with robust standard errors) relating atrophy rate to CAG repeat length, group and their interaction was used

to investigate whether the relationship between atrophy rate and CAG repeat length differed between groups, having adjusted for age at baseline. A similar model was used to assess the relationship between atrophy rate and UHDRS motor score. If there was no evidence of an interaction then the analysis was repeated without interaction terms, with and without adjusting for differences due to group, and adjusting for age.

11.2.3.3 Associations with change in clinical and cognitive measures

Separate linear regression models (adjusting for age) were used to assess the relationship between atrophy rate and change in UHDRS independence score and total functional capacity (in the HD group alone).

For each of the cognitive change variables a linear regression model (with robust standard errors) relating atrophy rate to change score, group and their interaction was used to investigate whether the relationship between atrophy rate and change score differed between groups, having adjusted for age at baseline and estimated premorbid IQ at baseline. Where it did, slopes in each group were examined to see whether they were significantly different to zero. If there was no evidence of an interaction then the analysis was repeated without interaction terms, but adjusting for differences due to group, age and IQ.

Where a statistically significant association between atrophy rate and cognitive decline was found, the above analysis was repeated adjusting for the effects of CAG repeat length by including it as a covariate.

11.2.3.4 Excluded scan pairs

In order to check whether scan-pair exclusion had biased the remaining sample Fisher's exact test was used to see whether the proportion of accepted scans differed between groups. Following this, Fisher's exact test was used to investigate differences in gender and stage between the accepted and rejected subjects in each group, and unpaired t-tests (two-tailed, allowing for unequal variances where necessary) were used to check for differences in other demographic and clinical variables.

11.3 RESULTS

Demographic data are shown in Table 11-1. There were no statistically significant differences in gender, handedness or estimated premorbid IQ between the groups. The mean age of the premanifest group was 8.7 years lower than controls (95% CI 2.8, 14.7 years, p=0.005). The premanifest group had a shorter mean CAG repeat length than the early HD group (mean difference 1.3 repeats, 95% CI 0.2, 2.3 repeats, p=0.023).

Mean (SD) interval between baseline and follow-up MRI scans was 1.0 (0.1) year for each group and did not differ significantly between groups.

Table 11-1 Demographic data for the longitudinal imaging cohort

	Control (N=18)	Premanifest (N=17)	Early HD (N=27)
Gender (M:F)	6:12	8:9	16:11
Age (year)[a]	46.3 (9.7)	37.5 (7.4)	48.9 (9.0)
Estimated premorbid IQ	106.8 (11.5)	103.2 (8.7)	105.0 (13.2)
Handedness (R:L)	17:1	16:1	25:2
CAG repeat length[b]	NA	42.4 (1.7), range 40 – 45	43.7 (1.7), range 40 – 47
Predicted years to onset[c]	NA	16.8 (6.6), range 9 – 31	NA
Disease duration (year)	NA	NA	4.5 (2.9)
UHDRS Motor[d]	1.1 (0.9)	3.9 (4.4)	28.1 (11.9)
UHDRS Independence	100 (0)	100 (0)	92.0 (9.1)
UHDRS TFC	13 (0)	13 (0)	11.1 (1.7)
Scan interval (year)	1.0 (0.1)	1.0 (0.1)	1.0 (0.1)

NA: not applicable

Data are mean (SD) with the exception of gender and handedness; handedness was taken as the hand used to write with; UHDRS: motor is out of 124, higher score = more severely impaired; independence is scored as a percentage, higher score = better function; Total Functional Capacity is out of 13, higher score = better function.

a PM<Control p=0.005, PM<HD p<0.0001

b PM<HD, p=0.023

c Calculated from age using the equation from Langbehn *et al.* (2004), with onset defined as a 60% chance of showing motor signs

d HD<PM, HD<Control, both p<0.0001

11.3.1 Application of the BBSI

Representative sagittal slices from scans from each timepoint for a control and early HD subject (matched for age and gender) are shown in Figure 11-1. The difference in caudate size is notable between the two subjects. The colour overlay shows regions of loss (red) or gain (green) between the two timepoints, detected in the boundary region by the BBSI algorithm. Although the main shift in boundary in the HD subject is around the caudate, there is also clear ventricular enlargement and some increase in the sulcal spaces. There is very little change in the control over the same period.

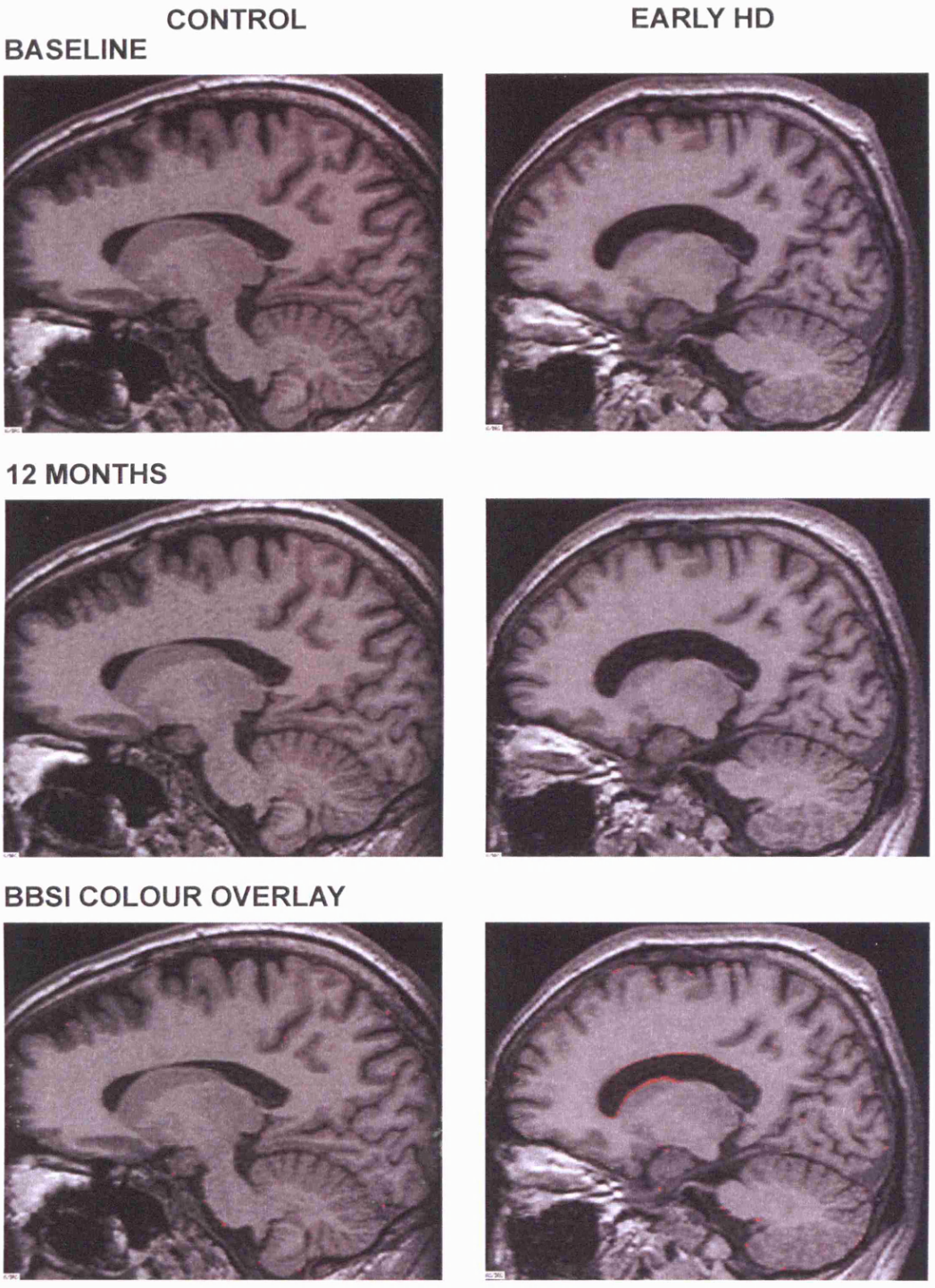


Figure 11-1 Sagittal sections through the left caudate of a control (left panel) and early HD subject (right panel) taken at baseline, 12 months, and showing the BBSI as a colour overlay

Subjects were male and aged 34 at baseline. The control atrophy rate was 0.2% (of baseline brain) per year, and the early HD subject 1.2% per year. Red shows regions of loss, green of gain (defined as a change of >20% of mean brain intensity).

11.3.2 Group differences

The HD group had a higher mean rate of atrophy compared with controls (0.92% per year vs. 0.38% per year, p=0.005) and the premanifest group (0.92% per year vs. -0.06% per year, p<0.0001) after adjusting for age (see Table 11-2 for means and Figure 11-2 for individual rates standardized to mean age in the whole cohort). Although mean atrophy rates were higher in controls than in the premanifest group this difference was not statistically significant after adjusting for age (p=0.10).

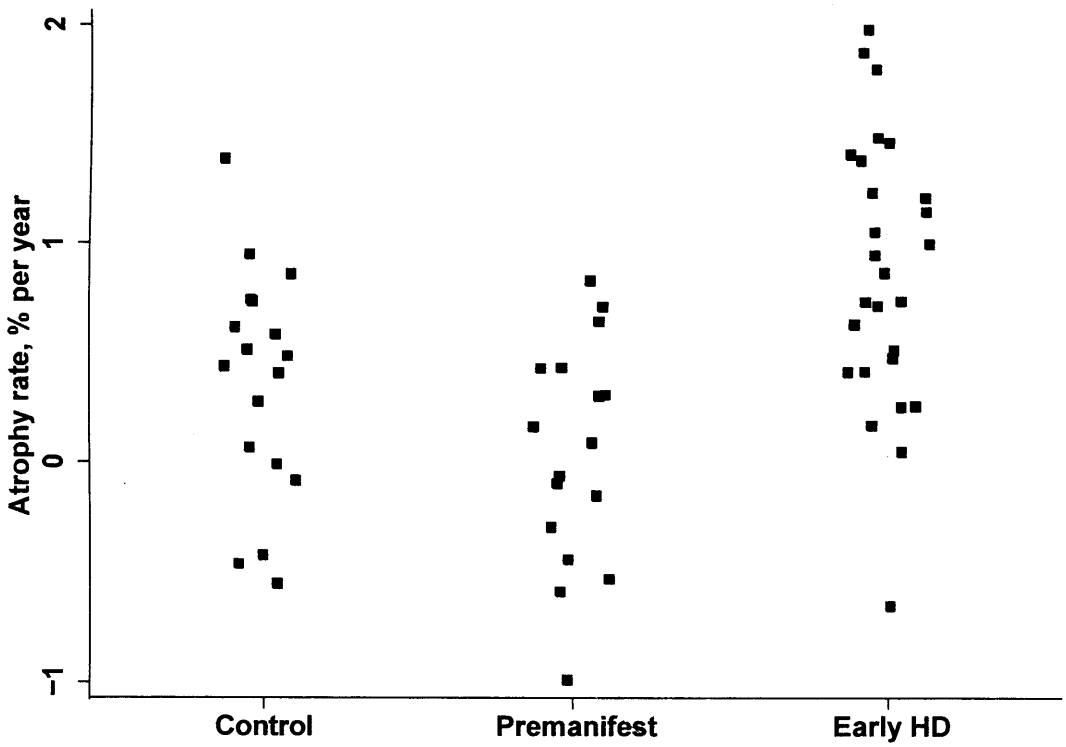


Figure 11-2 Whole-brain atrophy rates (% per year) in each group, standardized to mean age in the whole cohort

Table 11-2 Mean (SD) atrophy rates with differences with and without adjustment for age at baseline

	Control	Premanifest	Early HD	Difference (Premanifest - Control)		Difference (Early HD - Control)	
	N=18	N=17	N=27	Crude	Adjusted	Crude	Adjusted
Atrophy rate (%/yr)	0.38 (0.51)	-0.06 (0.47)	0.92 (0.67)	-0.44 (p=0.011)	-0.31 (p=0.10)	0.55 (p=0.003)	0.51 (p=0.005)
95% CI	0.12, 0.63	-0.30, 0.18	0.66, 1.19	-0.77, -0.10	-0.68, 0.06	0.19, 0.90	0.16, 0.85

Positive differences indicate a higher atrophy rate in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age at baseline; adjusted differences are the differences having adjusted for age by including it as a covariates in the linear regression model.

11.3.3 Associations with baseline clinical measures

In the early HD group there was no evidence of an association between atrophy rate and baseline disease duration, UHDRS independence score or TFC (all $p>0.43$). There was also no evidence of an association between atrophy rate and probability of onset in the premanifest group ($p=0.29$).

Combining subjects from both patient groups an increase of CAG repeat length by one was associated with an increase in whole-brain atrophy rate by 0.12% per year after adjusting for age (95% CI 0.005, 0.24% per year, $p=0.042$) (Figure 11-3).

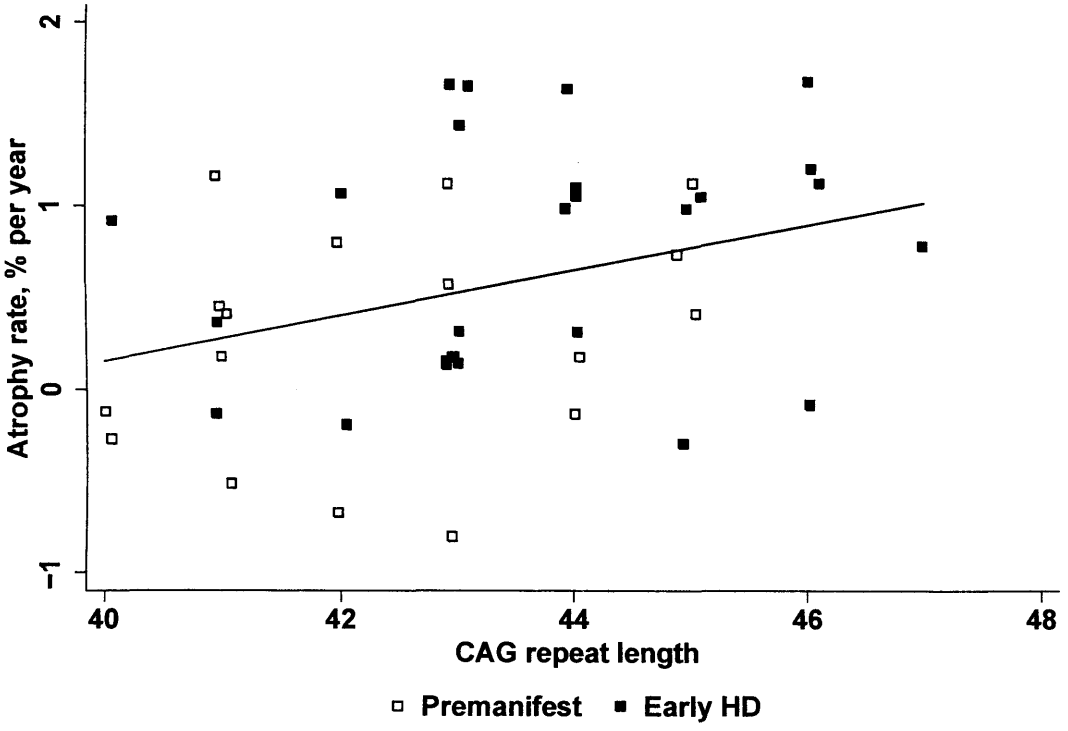


Figure 11-3 Atrophy rate (% per year, standardised to mean age in the whole cohort) plotted against CAG repeat length

This association reflects, in part, the fact that CAG repeat length is associated with earlier progression to manifest HD and that early HD subjects have higher atrophy rates than premanifest subjects. After adjusting for this, the slope was

close to zero, and not statistically significant (p=0.79). There was no evidence that the slope of this association differed between the patient groups (p=0.37).

After adjusting for age, there was no evidence that the association between baseline motor score and atrophy rate differed between premanifest and early HD subjects (p=0.51). In the gene carriers as a whole a one point increase in motor score was associated with a 0.02% per year increase in atrophy rate (95% CI 0.01, 0.04, p=0.005) but as above, this could be explained by group differences and the association was no longer statistically significant when the effect of group was adjusted for (p=0.42).

11.3.4 Associations with change in clinical and cognitive measures

In the early HD group there was no evidence of an association between whole-brain atrophy rate and decline in UHDRS independence score or TFC (both p>0.76).

In the gene carriers as a whole higher whole-brain atrophy rate was associated with greater decline at Stroop word reading (-4.3 points per 1% increase in rate, 95% CI -8.3, -0.4 points, p=0.034) and TMT B (-17.5 seconds per 1% increase in rate, 95% CI -32.6, -2.4 seconds, p=0.024), after adjusting for age and IQ. Higher whole-brain atrophy rate was also associated with less decline in digit span backwards in the early HD group, but not the premanifest group (1.7 points per 1% increase in rate, 95% CI 0.1, 3.3 points, p=0.039 in the early HD group). Atrophy rates were not significantly associated with change in any of the other cognitive variables.

When CAG repeat length was included in this analysis higher atrophy rates were still associated with greater decline in Stroop word reading (-4.2 points per 1% increase in rate, 95% CI -8.1, -0.3 points, p=0.035) and TMT B score (-17.3 seconds per 1% increase in rate, 95% CI -33.1, -1.5 seconds, p=0.033) in the gene carriers as a whole. The relationship between digit span backwards and atrophy rate in the early HD group was also unchanged.

11.3.5 Excluded scan pairs

Demographic data for included and excluded subjects are shown in Table 11-3. There was a significant difference in the proportions of subjects from each group who were rejected, with no controls (0%), 2 premanifest (11%) and 9 early HD subjects (25%) rejected (p=0.034). Within the premanifest group there were no statistically significant differences between the accepted and rejected subjects for any demographic or clinical variables at baseline or follow-up except for CAG repeat length (lower in rejected subjects, p=0.0008) and baseline motor score (lower in rejected subjects, p=0.016). Within the early HD group there were again no differences between the accepted and rejected subjects for most demographic and clinical measures, except that a higher proportion of women than men were rejected (p=0.002), and rejected subjects were more depressed at follow-up than accepted ones (p=0.0254).

Table 11-3 Demographic data for the included and excluded subjects in each patient group

	Premanifest		Early HD	
	Included (N=17)	Excluded (N=2)	Included (N=27)	Excluded (N=9)
Gender (M:F)	8:9	1:1	16:11	0:9*
Age (year)	37.5 (7.4)	39.5 (12.0)	48.9 (9.0)	47.8 (12.6)
Stage (1:2)	NA	NA	20:7	4:5
CAG repeat length	42.4 (1.7)	40.0 (0)*	43.7 (1.7)	43.6 (3.8)
Predicted years to onset	17.0 (10.3)	26.5 (11.4)	NA	NA
Disease duration (year)	NA	NA	4.5 (2.9)	3.6 (2.1)
UHDRS Motor (baseline)	3.9 (4.4)	1.0 (0)*	28.1 (11.9)	31.0 (14.5)
UHDRS Motor (12 months)	5.4 (4.5)	4.0 (2.8)	34.3 (12.2)	34.1 (13.7)
UHDRS Independence (baseline)	NA	NA	92.0 (9.1)	88.3 (11.7)
UHDRS Independence (12 months)	NA	NA	89.4 (9.0)	88.3 (11.7)

UHDRS TFC (baseline)	NA	NA	11.1 (1.7)	10.4 (2.0)
UHDRS TFC (12 months)	NA	NA	10.4 (1.9)	10.6 (1.9)
BDI (baseline)	5.8 (6.1)	9.0 (2.8)	7.8 (6.9)	13.8 (13.1)
BDI (12 months)	6.5 (7.6)	11.5 (4.9)	7.3 (6.8)	14.1 (9.9)*

* excluded statistically significantly different to included, $p<0.05$

11.4 DISCUSSION

This work demonstrates that rates of whole-brain atrophy are increased in early HD, and that this can be detected with serial MRI and the BBSI over one year. The mean rate of whole-brain atrophy, of just under 1% per year in the HD group, was similar to that found previously, in a smaller sample examined over a six-month period, reported in chapter 3.

There is increasing evidence that neurodegeneration in HD occurs throughout the brain, even in the early stages (Beglinger *et al.* 2005; Rosas *et al.* 2005; Paulsen *et al.* 2006b) and this is reinforced by these findings. A whole-brain rate of 1% per year equates to about 10ml of loss per year, and the total volume of the healthy striatum is only approximately 20ml (Aylward *et al.* 2004); therefore striatal atrophy alone, occurring at rates of <5% per year in manifest HD subjects (Aylward *et al.* 2003), cannot account for the whole-brain rates seen here. The control atrophy rate, of 0.38% per year, was also similar to previously reported figures (Scahill *et al.* 2003). This cohort of manifest HD subjects consisted of those in the early stages of the disease, with an average range of CAG repeat lengths (40–47), and was thus representative of the relatively mildly affected and young HD population most likely to go into clinical trials of disease-modifying treatments.

These findings differ from an earlier study which found evidence that white matter (but not grey matter) volumes decreased over time in premanifest subjects, but that neither grey nor white matter volume decreased in early HD subjects (Ciarmiello *et al.* 2006). However the segmentation algorithm used there was fully automated with reproducibility of around 1.1% (calculated as the

standard deviation of repeated measures as a proportion of their mean); the equivalent figure for our semi-automated manual method is ~0.3% for both intra- and inter-rater reproducibility. In addition, the BBSI uses intensity shifts to quantify volume change in a region defined by the segmentations, rather than segmentations *per se*, and so is less susceptible to segmentor error than methods which rely on segmentation alone as an estimate of volume. It is likely that relative to Ciarmiello *et al.*'s method, the BBSI is more sensitive for detecting change on the scale seen here, which is small relative to some other neurodegenerative diseases (Fox and Freeborough 1997).

There was no evidence of increased atrophy rates in the premanifest group but the majority of this group was more than 15 years from predicted motor onset, and there was also no evidence that this group differed from controls on most cross-sectional measures, (see chapter 4). However, this group did have significantly smaller whole-brain volume than controls, which could be due to either a developmental abnormality, or very slightly increased atrophy rates acting over several years, or both. Others have suggested abnormal gyrification in premanifest subjects, as well as increased grey (but not white) matter (Paulsen *et al.* 2006b; Nopoulos *et al.* 2007), in the absence of differences in total brain size. Further investigation is therefore needed to determine the relative contribution of these two possibilities to the volumetric findings presented here. Given the findings of Paulsen *et al.* (2006b), one would need to posit slightly increased atrophy rates in order to explain the decreased brain and caudate volumes in this premanifest cohort. However, the BBSI technique may not be sensitive enough to pick up changes in whole-brain atrophy rates during these very early stages. This result may also suggest that cerebral volume losses are

likely to be subtle, slow and possibly regionally confined many years before onset (Aylward *et al.* 2004; Paulsen *et al.* 2006b).

An increase of CAG repeat length by one CAG was associated with an increase in whole-brain atrophy rate of 0.12% per year in the entire HD gene carrier sample, with CAG repeat length accounting for about 10% of the variance in atrophy rate. This was largely explained by the fact that people with a higher number of CAG repeats tended to be those with manifest disease, and also those with higher rates of atrophy. Although in other domains such as neurological function and cognition, rate of decline has been reported to increase with disease duration or severity (Ward *et al.* 2006; Rosenblatt *et al.* 2006), that was not the case for whole-brain atrophy rates, which did not vary significantly with proximity to motor onset (in the premanifest group), duration of motor signs (in the HD group) or putative measures of disease severity such as motor score, TFC, or clinical stage. Whole-brain atrophy rate may be relatively constant, at least over the period of early disease examined here. It is also likely that motor signs are more strongly associated with regional, rather than global atrophy, particularly that of the putamen and associated motor areas in the frontal cortex.

However, although whole-brain atrophy rates were not predictive of motor decline, higher atrophy rates were associated with greater decline in some cognitive tests (both before and after taking into account the effect of CAG repeat length), notably timed psychomotor tasks tapping both automatic skills and cognitive switching. There is evidence that such skills are associated with cortical as well as sub-cortical regions (Rosas *et al.* 2005; Kassubek *et al.* 2005; Hirshorn and Thompson-Schill 2006) all of which the BBSI, as a whole-brain

measure, encompasses. In the early HD group, higher atrophy rate was also associated with less decline on digit span backwards although it would be unlikely for this to represent a causal association. The robustness of these associations will be followed up in a subset of subjects who are having a third assessment approximately 27 months after baseline, in order to assess whether the changes detected at one year are representative of those measured over a longer period.

The BBSI is relatively fast, unbiased and operator efficient (two whole-brain images can be processed in ~30 minutes); registration corrects for voxel size changes introduced by scanner drift; and the impact of segmentor variability is limited because the technique measures the intensity shift in the images, and does not rely on the segmentations *per se* as a measure of volume. Furthermore the scanning time is relatively short and well-tolerated, with ninety-five percent of our initial cohort returning for the follow-up assessment.

Some registered scan pairs were not analysed because of scanning artefacts (e.g. flow artefact affecting temporal lobe signal). Subjects whose images were rejected did not have higher motor scores than those who were accepted, nor were they more advanced on any other measure, suggesting that non-disease-related factors influenced scan quality. However, in the case of those subjects who were unable to produce useable images in the first place, this was due to disease-related factors such as chorea or claustrophobia. Any imaging technique is likely to be subject to such drawbacks, particularly in this patient population, and MRI measures may not be suitable as biomarkers in the later stages of the disease when motor signs become more severe.

Further work should add to this finding by investigating which brain regions (other than the striatum) are contributing to the increased whole-brain atrophy rates seen here. Techniques such as cortical thickness measurements, and voxel-based morphometry have the potential to localise regions of greater or smaller brain volume loss over time, although as yet they have rarely been applied longitudinally in HD, and where they have, techniques differ widely (e.g., Kipps *et al.* 2005; Ruocco *et al.* 2007). However, increased knowledge of the focal regions of atrophy in HD, which may vary with disease stage, might suggest a number of potential structural biomarkers appropriate for use in different HD populations.

11.5 CONCLUSION

Overall whole-brain atrophy rate meets some of the criteria required for a biomarker (Biomarkers Definitions Working Group 2001; Aylward 2007). It can be measured objectively, is associated with the length of the abnormal CAG repeat that causes HD and predicts some cognitive decline. Importantly, in encompassing the whole brain, the BBSI is capturing the full extent of the diffuse cerebral volume loss in HD and hence might be relatively more sensitive to disease modification by treatments that target multiple brain regions, than would a more focal measure. There was no evidence that whole-brain atrophy rate was increased in the far-from-onset premanifest group, whereas it has been shown in a premanifest population closer to predicted onset that caudate atrophy rates are increased and predictive of motor onset (Aylward 2007). Ultimately it is likely that different markers, or combinations of markers, will be appropriate for different disease stages, and that practical considerations including necessary sample size will also influence the choice of marker for a particular trial. Based

on the findings in this study 278 early stage HD subjects per treatment arm would need to complete a one-year trial in order to detect a 20% reduction in whole-brain atrophy rate (note that this is a reduction in total atrophy rate, not relative to controls) (Fox *et al.* 2000). The BBSI technique is already used as a marker in other neurodegenerative diseases (Fox *et al.* 2005), and would seem to be an attractive measure for clinical trials in the early symptomatic stages of HD. Further work is now needed, particularly in the premanifest population, to see at what point an increase in whole-brain atrophy becomes detectable and whether rates do remain constant through early disease.

12 CONCLUSIONS

12.1 CLINICAL FINDINGS

This thesis used serial MRI and cognitive testing to investigate the cognitive deficits and brain volume loss seen in premanifest and early Huntington's disease and to assess the potential of these measures to detect change over time.

Chapter 3 demonstrated that serial MRI is well-tolerated in HD, that the BBSI can be applied to measure atrophy in this cohort and that whole-brain atrophy rates are increased in early HD.

Chapter 4 showed that whole-brain volume was already decreased at baseline assessment, in both early and relatively far-from-onset premanifest subjects. Cognitive deficits were fairly widespread in the early HD group, although of particular note was the finding that deficits in recognition memory and facial emotion recognition were of similar magnitude to those in psychomotor and executive function tasks. Recognition of true target stimuli was relatively good, given the level of encoding achieved by early HD subjects, confirming previous work. However there was a tendency for these subjects to make more false positives, both confirming theories that storage *per se* is not the main domain of impairment, and suggesting that strategies for monitoring retrieval, or making confidence judgements, might be impaired. Although naming ability was unimpaired, naming latency was slowed, and this was not explained by dysarthria, suggesting that cognitive as well as motor slowing was contributing to the impairments in this group.

As has been found in a number of other studies, there was little evidence of cognitive impairment in the premanifest group, which was relatively far from predicted motor onset. Interestingly the one task on which these subjects were slower than controls was pressing a button in a predictable motor task, which seems indicative of early striatal damage given the proposed role of the striatum in goal-directed behaviour and habit learning.

Caudate volume was also decreased in both patient groups (chapter 5) further suggesting that striatal atrophy could be present even in those very far from estimated disease onset, and had probably been ongoing for some time prior to entry to the study. Together with the whole-brain findings these results imply an insidious and early increase in cerebral atrophy, initially focused on striatal regions. Although research on caudate laterality has been equivocal, the work presented here added to previous suggestions of a slight normal rightward asymmetry, and this was found in both patients and controls. There was an interesting dichotomy in the associations between cognitive performance and caudate volume, that supported the role of the striatum in learnt (automatic) skill performance, goal-directed behaviour and in motor, oculomotor and psychomotor circuits, but suggesting that striatal damage might not underlie deficits seen in memory tasks, or tasks under more cognitive control, such as those involving a switching component.

In terms of the locus of atrophy across the whole brain, chapter 7 demonstrated that it was more widespread than previously thought, including the insula, and regions of the frontal, parietal and occipital lobes, as well as subcortical regions. Higher CAG repeat length was associated with increased atrophy of the striatum

as well as some extra-striatal regions. Chapter 8 went on to look at emotion recognition in more detail, showing a broad impairment of negative emotions with anger and fear at least as badly affected as disgust. These impairments were all associated with reduced grey matter in the striatum predominantly, as well as in some other regions, suggesting that they might be caused by damage to a common underlying network, most likely including fronto-subcortical circuits. This did not support theories that atrophy in HD causes selective deficits to recognition of specific emotions (namely disgust), although further work, probably including fMRI, is needed to confirm the underlying neural substrates of the impairments.

12.2 TRACKING CHANGE

Whole-brain atrophy rates were significantly increased in early HD and this increased loss was detectable over only one year, using serial MRI and the BBSI (chapter 11). Atrophy rates were variable, with some overlap between groups, and this was in part explained by CAG repeat length, but not by disease duration or the interaction between CAG repeat length and disease duration. Given that HD is clearly a whole-brain disease, even in the early stages, this semi-automated, robust technique shows promise. Further work will assess the effect of longer intervals between scans, whether other variables contribute to the variance, and whether the BBSI is also sensitive to change in the close-to-onset premanifest population. It will also be important to assess other measures of atrophy, such as cortical thickness, and other imaging modalities, for example using diffusion imaging to assess white matter deficits.

Cognitive change over one year was generally small and variable (chapter 9), although in the early HD group decline was detected in both executive function and memory tasks. Lack of significant decline in other tasks might be due to slowness of progression or because this group had already undergone decline and was performing relatively poorly at both timepoints; in both cases the relative insensitivity of tests could also be a factor, and this could be addressed, particularly in the memory domain. The premanifest group showed a relative decline in phonemic fluency, and although this was only in relation to the large practice effect seen in controls, it nevertheless suggests some early prefrontal dysfunction which could be a focus of further testing in this group.

CAG repeat length was associated with smaller whole-brain and caudate volume, as well as higher whole-brain atrophy rates. VBM analysis suggested that it was also related to extra-striatal volume although this was quite focal (chapter 7). There was little evidence that it was associated with greater rate of cognitive or clinical decline, however, although this may be due to the variability of cognitive and clinical change. It may also reflect the fact that the subjects studied here had CAG repeat lengths of between 40 and 50, i.e. in the low pathological range, and it may be that the effect of CAG repeat length on rate of decline becomes more apparent in subjects with higher repeat lengths. In addition, very few of the cognitive, clinical or MRI measures employed here showed rate of change to be associated with disease duration.

12.3 TECHNICAL CONSIDERATIONS

With increasingly diverse imaging techniques in use, many differences in findings between studies may be attributed to these, rather than true differences

in the populations studied. Confounding factors need to be considered for both manual and voxel-based morphometric techniques, as was demonstrated in chapter 6, in particular the effects of head size and gender. Adjusting for total intracranial volume corrects for differences due to gender, but it is important to note that this is not the case for all estimates of head size (for example using intracranial area instead of volume) and that some estimates may be biased (for example intracranial volume estimated by automated segmentation methods). Manual techniques can limit sources of bias by blinding investigators to laterality, and using the same segmentation protocol for each hemisphere. With automated techniques, such as SPM, is it vital to investigate the accuracy of any preprocessing steps for the particular cohort under investigation, and to consider these and other potential methodological problems in interpreting, and generalising results (chapter 6). With the advent of improved normalisation and segmentation techniques, such as DARTEL (implemented in SPM5, (Ashburner 2007)) it should be possible to reduce some of the errors and hence findings will be more robust. This work also suggested and compared methods for ensuring maximum agreement between the BBSI and "true" volume loss (chapter 10).

Cognitive testing can be unreliable because of the many non-disease-specific factors that can affect performance, including demographic variables such as age and IQ, which can have small (but potentially significant) effects on findings. As with imaging methods, the effects of these confounding variables can be adjusted for statistically, in order to minimise their impact on findings.

In this thesis cognitive findings were not adjusted for multiple comparisons, because each test was considered of independent interest, and therefore the risk

of missing true positives was kept low. Whether or not researchers opt to control the false positive rate more stringently it would be helpful if all studies reported mean effects, since in this way it is possible to compare results, even if some smaller studies have not found statistically significant differences (perhaps due to lack of power). Choice of control group is also important, and it may be possible to minimise the effects of some non-disease-related factors by recruiting controls from HD families, who have similar environmental, social and emotional influences.

In general, particularly with the publication of guidelines as to the reporting of some widely-used imaging techniques, it is to be hoped that future work increasingly confirms what is known about HD, and generates new hypotheses, rather than queries about methodology.

12.4 LIMITATIONS

The early HD cohort studied here was relatively large, and all subjects were stage 1 or 2 and their HD was of fairly short duration, meaning that the group was not too heterogeneous. However the premanifest group was smaller, and on average very far from predicted onset. Thus, whilst results might generalise to a similar population, this cohort needs continued follow-up in order further to clarify the nature of the deficit in premanifest subjects, to investigate when other early signs can be detected, and to confirm actual years to onset.

A time period of one year, whilst preferable for clinical trials, means that the signal was small and it would undoubtedly be improved with measurements taken over a longer period. It remains to be seen whether the benefits of this

outweigh the costs in the context of a trial for a potentially disease-modifying therapy.

There is currently some debate about the extent to which both CAG repeat length and disease duration affect rate of progression in HD, as well as prediction of motor onset in the premanifest population (see section 1.4). Although very little of the decline shown in this study (cognitive, clinical and atrophy rates) was associated with disease duration, it is likely that disease duration is a very imprecise measure; firstly the diagnosis of onset depends on the diagnosing clinician, and then how soon after "true" onset the patient is seen by that clinician; secondly, onset is still defined purely in terms of motor signs, although there is increasing evidence that behavioural and cognitive impairments may be present prior to this, and that the burden of different symptom types might also differ between patients.

There are similar problems with estimated years to onset. In the published literature a number of different methods have been used to estimate time to onset, meaning that seemingly similar groups of patients may in fact not be. Recently more researchers are using the equation developed by Langbehn *et al.* (2004), but this estimate seems somewhat conservative, based as it is usually on estimated time to reach a 60% probability of showing motor signs. Whilst classifying subjects in this way facilitates comparison between studies, it does not allow one to consider premanifest and early HD subjects on the same scale. Using this equation the predicted mean age at onset of the premanifest group presented here is 55.4 years, whereas the mean age at onset of the early HD group was 44.4 years, and although the premanifest subjects have, on average, slightly fewer

CAG repeats this seems unlikely to translate into such a large discrepancy in onset ages. Future work to refine the equation to take into account other potential predictors (such as the CAG repeat length on the normal gene, paternal or maternal inheritance, parental age at onset and other potential genetic modifiers) may give better estimates, as will follow-up of the current premanifest population to see whether there is a discrepancy between the real and predicted points at which they phenoconvert. Future work is also needed to assess the predictive (possibly additive) value of non-genetic measures (e.g. caudate volume or cognitive deficits). Finally, there is a move towards more consistent documentation of the earliest signs of the disease, whether they be motor, cognitive or behavioural, in an attempt to define "onset" (and thus disease duration) more accurately. Advances in these measures will be needed in order to draw more robust conclusions about the progression of the disease. It seems clear that different measures may be appropriate to track change at different stages of the disease process, and stratification accuracy (based on disease duration or years to onset) will need to improve in order to reduce within-group inhomogeneity and maximise the chances of detecting the optimum markers.

12.5 FUTURE WORK

The data and results presented in this thesis are encouraging, particularly in the use of whole-brain measures (such as the BBSI) to monitor change in HD, and the findings of specific types of cognitive impairment and decline over time. Further work will focus on expanding these findings: applying the BBSI to new cohorts, over longer time periods, and using the cognitive data to inform the development of more sensitive tests.

Memory tests may have great potential in HD because their difficulty can be manipulated. Controls and premanifest subjects tended to score at ceiling on some parts of both auditory and visual recall and recognition memory tasks, whilst early HD subjects were impaired to the extent that there was little scope to detect change on at least one task (some subjects were scoring towards floor levels on recognition memory for faces). These deficits seemed to be caused by poor initial encoding, and problems with recognition judgements in forced choice procedures, and there should be an interim stage (in the group of close-to-motor-onset subjects which was not well-represented in the current work) during which these deficits start to become detectable. Memory tasks can be manipulated in a number of ways: briefer stimulus presentation, longer time between encoding and recall, more items to be remembered, more distractor stimuli in recognition paradigms, and therefore it seems likely that tests could be developed which were more sensitive to early (frontal) dysfunction in HD, with the aim of adding to the very few tests that can detect deficits, if not decline, in the premanifest population.

The finding that premanifest subjects were slowed at a predictable button press also merits follow-up. This suggests that even in far-from-onset subjects, tasks that tap skill-learning, habit acquisition or automation of skills (all of which are thought to rely strongly on striatal integrity) are likely to be sensitive to some of the earliest signs of the disease. Future work will build on this, following-up the current cohort and investigating other potentially sensitive tasks, similar to the one used here.

The contribution of the eye movement impairments in HD to the deficits seen here is as yet unknown. Given that fear recognition impairment has, in one case, been attributed to erroneous gaze direction, and temporarily corrected by explicit instructions as to which part of the face to attend (Adolphs *et al.* 2005), it would be interesting to investigate the role of eye movements to the facial emotion recognition impairment in HD in particular, with a view to determining whether instructions could similarly improve recognition performance. Whilst it would be unlikely to improve performance to control levels, minimising the effect of poor scanning might be of some benefit to patients and their carers in their daily social interactions. Many of the cohort presented here are also undergoing assessment using a multimodal battery, including facial, vocal and musical emotion stimuli, in order to investigate whether the impairments seen here are modality-specific, and whether performance is associated with similar neural substrates.

All the subjects presented here have been invited back for further follow-up. Many are now enrolled in Track-HD (a three-year prospective study investigating a number of potential markers in premanifest and early HD). Those who are not (mainly premanifest subjects very far from predicted motor onset) continue to be offered annual assessments as part of the present study, and it will be vital to follow-up the premanifest subjects in particular, as when they phenoconvert it will be possible to reanalyse these data with accurate years-to-onset figures, and thus to assess the predictive power of a number of the tests used here.

12.6 SUMMARY

This thesis examined cognitive and MRI measures in a large cohort of premanifest and early Huntington's disease subjects over one year. Cross-sectionally it is clear that differences are apparent on MRI and cognitive testing, even in relatively far-from-onset premanifest subjects. Subjects tolerated serial MRI well, and whole-brain atrophy rate measured with the BBSI could potentially be a marker of progression. Despite the problems associated with repeated cognitive testing and the slow progression of the disease, some cognitive tests also showed promise as potential markers of change.

There are a number of conflicting findings in the HD literature and there have been calls for better-validated and consistent research in order to measure and track progression at a time when markers are vital if potentially disease-modifying therapies are to be assessed in clinical trials (Nopoulos *et al.* 2007; Witjes-Ane *et al.* 2007). In a large part these differences can be explained by methodological inconsistencies, and demographic (as opposed to disease-specific) differences in the populations being studied. The work in this thesis addressed many of these factors by identifying and adjusting for confounding variables, eliminating sources of bias, and validating imaging techniques. Only when techniques (both cognitive and imaging) are well-validated, as free from bias as possible, and after adjusting for any possible confounds, will it be possible in future to draw firm conclusions and move on to new research questions.

This work adds to the body of knowledge showing that atrophy in HD is not confined to the striatum, and that cognitive deficits are wide-ranging and in some

cases detectable many years before motor signs are apparent. Although motor slowness and chorea are tangible signs of the disease, deficits in skill learning, phonemic fluency, memory, and emotion recognition are also apparent early in the disease, and in some cases prior to motor onset. Future work should aim for better understanding, and hence management, of these very early symptoms, in order to help those patients, families and carers who live with the devastating effects of this disease.

PUBLICATIONS AND COLLABORATIONS

Peer-reviewed papers based on results described in this thesis are detailed below together with the contributions of all other individuals who were involved in this work. I am indebted to all of my colleagues for these contributions.

Pilot study (Chapter 3)

Henley, S.M.D., Frost, C., MacManus, D.G., Warner, T.T., Fox, N.C., Tabrizi, S.J. (2006). Increased rate of whole-brain atrophy over 6 months in early Huntington's disease. *Neurology*, 67, 694-696.

Subjects were recruited by me. Sarah Tabrizi and Tom Warner assessed patients using the UHDRS. Neuropsychological assessment was carried out by me under the supervision of Elizabeth Warrington, and volumetric analysis was done by me. Registrations were checked by Rachael Scahill. Statistical advice was provided by Chris Frost. I carried out the analysis and drafted the paper which was revised by Chris Frost, Sarah Tabrizi and Nick Fox.

London longitudinal study (Chapters 4 – 11)

Subjects were recruited by me, Edward Wild and Roger Barker. Neuropsychological assessment was administered by me, with advice from Elizabeth Warrington and Seb Crutch. The BDI and clinical parts of the UHDRS were administered by Edward Wild. Volumetric measurements and registrations were done by me and Nicola Hobbs (whole brains, caudates and TIVs). Images were quality-checked by Shona Clegg. Statistical advice was provided by Chris Frost and technical VBM advice by Ged Ridgway and Rachael Scahill. All

analysis and voxel-based morphometry was performed by me. Further details specific to each chapter are listed below.

Chapter 7

Henley, S.M.D., Wild, E.J., Hobbs, N.Z., Scahill, R.I., Ridgway, G.R., MacManus, D.G., Barker, R.A., Fox, N.C., Tabrizi, S.J. Relationship between CAG repeat length and brain volume in premanifest and early Huntington's disease. *Journal of Neurology* (in press)

Recruitment, assessment and analysis were carried out as outlined above. I drafted the paper which was revised by Ged Ridgway, Roger Barker, Sarah Tabrizi and Nick Fox.

Chapter 8

Henley, S.M.D., Wild, E.J., Hobbs, N.Z., Warren, J.D., Frost, C., Scahill, R.I., Ridgway, G.R., MacManus, D.G., Barker, R.A., Fox, N.C., Tabrizi, S.J. 2008. Defective emotion recognition in early HD is neuropsychologically and anatomically generic. *Neuropsychologia*, 46, 2152-2160.

Recruitment, assessment and analysis were carried out as outlined above. Jason Warren and Rohani Omar provided useful advice about creating ROIs in MRIcro and Jo Foster provided an amygdala segmentation on which to base one of the ROIs. I drafted the paper which was revised by Jason Warren, Ged Ridgway, Roger Barker, Sarah Tabrizi and Nick Fox.

Chapter 11

Henley, S.M.D., Wild, E.J., Hobbs, N.Z., Frost, C., MacManus, D.G., Barker, R.A., Fox, N.C., Tabrizi, S.J. Whole-brain atrophy as a biomarker of progression in early Huntington's disease. *(submitted to Movement Disorders)*

Recruitment, assessment and analysis were carried out as outlined above. I drafted the paper which was revised Roger Barker, Edward Wild, Sarah Tabrizi and Nick Fox.

ACKNOWLEDGEMENTS

"How shall I ever find the grains of truth embedded in all this mass of paper?"
Virginia Woolf, A Room of One's Own

On starting this work I felt as one great writer did, faced with the might of the British Library (or in my case PubMed). I am immensely grateful to everyone mentioned below for their help and support.

Firstly, and most importantly, this PhD could never have existed without my willing subjects and their lovely families, friends and carers, all of whose courage, enthusiasm and support never fail to move me. It is a privilege to work with you all.

The majority of this work was funded by the High Q Foundation and I would particularly like to thank Allan Tobin and Ethan Signer for their encouragement. Roger Barker helped enormously by asking members of his clinic to participate, and his manuscript revisions turned many a rough draft into something much better. Dave Mac is owed many pints thanks to his patience, willingness to scan at all hours and on all days of the week, and unfailing ability to get decent images from nervous subjects.

Nothing would have been significant without the expert help of Chris Frost, to whom I owe a huge debt for his patience in the face of countless questions, and willingness to explain everything algebraically. There seemed to be no MATLAB or VBM problem too insurmountable for Ged Ridgway who gave much of his time and advice to help and was invaluable for getting the blobs in the right places. Stefan Klöppel and Jason Warren were always on hand for discussion and advice, and much of this work would have been the poorer without them.

To all the DRC imagers, psychologists, programmers, nurses, clinicians and admin staff who make coming in to work such a pleasure I owe my thanks; some people in particular deserve a mention: Nicky, a joy to work with, thanks for everything, especially all the late nights, caudates, dolphin impressions; Shona,

countless cups of tea and QC; Jo B for making me laugh and lovely pictures of brains; Val for being calm and sharing PhD woes; Ali for trying to keep the server going and Rich for mending my computer when it failed; Seb and Elizabeth for words of wisdom on all things psychological; Julia, fellow musician and fantastic desk-mate; Riitta, Katy and CANDID for putting everything in perspective; Ed for his contribution and all the other clinicians for keeping cake in the office; Gill for guiding me (and Nick) through the application process; Clare, Suzie and Anne for being voices of calm in the midst of (occasional) chaos.

To the HD clinic staff and team who do such a marvellous job, thanks for all your support and help with recruiting, especially Rachel Taylor, Marianne Novak and the Track staff. Thanks also to my friends and colleagues in Euro-HD and Track-HD, in particular Jen Thompson and Julie Stout, who have all willingly shared thoughts and ideas.

To Nick, Sarah and Rach, a triumvirate that's hard to beat: I'm indebted to you for all that I've learnt and in particular to Nick, for taking a chance on me and turning me into a psychologist who knows about imaging; Sarah, for your unfailing support and energy, and your knowledge of all things (and people) HD-related; and Rach, for being a superb mentor, and friend, throughout your time at the DRC and beyond.

Finally to all my other friends who always said I would get there in the end, thank you. And to my family, and my lovely lad, your support means more to me than anything, and I couldn't have done it without you.

APPENDIX 1: IMAGE ACQUISITION PROTOCOLS

Pilot study (Chapter 3)

Images were obtained using an IR prepared FAST spoiled GRASS sequence with 24cm x 75% field of view and 256 x 256 matrix, reconstructed to provide 124 contiguous 1.5mm-thick coronal slices. In-plane pixel dimensions were 0.9375 x 0.9375mm. Acquisition parameters were: repetition time = 15ms; echo time = 5.4 ms; flip angle = 15°; inversion time = 650ms; receiver bandwidth = 16kHz, NEX=1.

All other studies (Chapters 4 - 8)

Images were obtained using an IR prepared FAST spoiled GRASS sequence with a 24cm x 75% field of view and 256 x 256 matrix reconstructed to provide 124 contiguous 1.5-mm thick coronal slices. In-plane pixel dimensions: 0.9375 by 0.9375 mm. Acquisition parameters: repetition time = 13ms; echo time = 5.2ms; flip angle = 13°; inversion time = 650ms; receiver bandwidth = 16kHz, NEX=1.

APPENDIX 2: VOLUMETRIC ANALYSIS

Volumetric analysis was performed using MIDAS software (section 2.5.3.1) following the protocols detailed below. All segmentations were done on T1-weighted images. Reproducibility error was estimated as the mean absolute percentage difference between repeat segmentations on the same scan, and with the intra-class correlation coefficient.

Whole brain

Whole-brain segmentation was done using the protocol described by Freeborough *et al.* (1997). This algorithm is semi-automated and identifies brain voxels using interactive thresholding. The region thus outlined is eroded and then dilated a number of times to isolate non-brain tissue such as scalp and dura (see Figure A2-1). All segmentations for the pilot study (chapter 3) were performed by me, and those for the other studies by me and Nicola Hobbs. Intra-rater reproducibility (intraclass correlation coefficient) was >0.99 for both segmentors (SH, average difference 0.19% of mean; NH, average difference 0.27% of mean); inter-rater reproducibility was >0.99 (average difference 0.52% of mean).

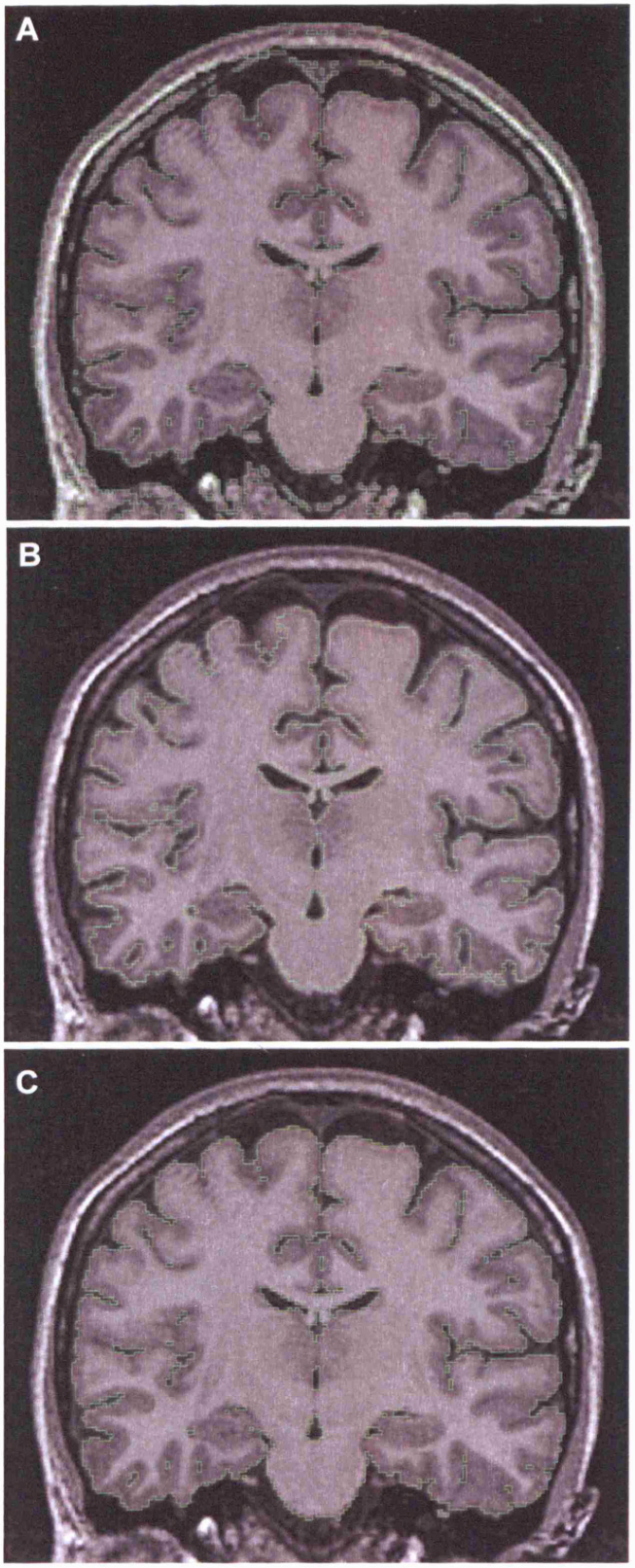


Figure A2-1 A) MRI showing intensity thresholding excluding high and low (non-brain) intensities; B) the region has been eroded by one voxel (cutting off skull); C) the region has been dilated by two voxels, expanding to the edges of the brain

Total Intracranial Volume

Total Intracranial Volume (TIV) was delineated using a protocol described by Whitwell *et al.* (2001). A lower threshold of 30% of mean brain intensity was set to outline the edge of the dura, which was then edited manually where necessary. The inferior boundary was the lowest slice in which cerebellar tissue could be seen. Every tenth axial slice was segmented with the superior boundary being the highest slice in which brain tissue could still be seen (see Figure A2-2). A volume estimate was obtained using linear interpolation. All measurements for the pilot study (chapter 3) were performed by me, and those for the other studies by me and Nicola Hobbs. Intra-rater reproducibility was >0.99 for both segmentors (SH, average difference 0.07% of mean; NH, average difference 0.56% of mean); inter-rater reproducibility was >0.99 (average difference 0.24% of mean).

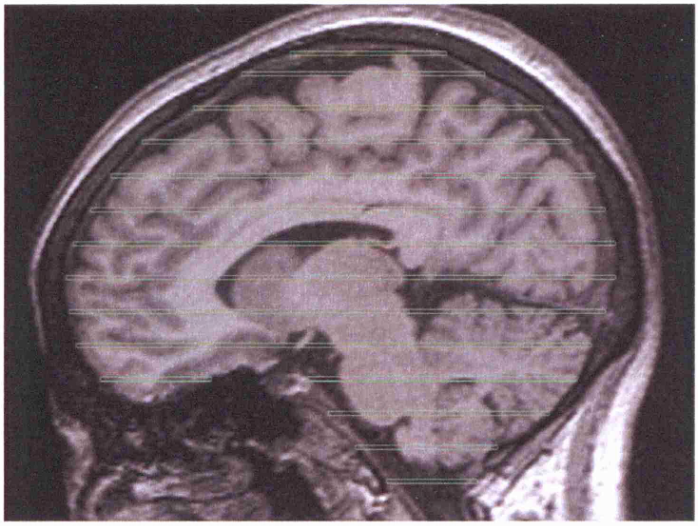


Figure A2-2 TIV segmentation in sagittal view

APPENDIX 3: NEUROPSYCHOLOGICAL ASSESSMENT

The following standard neuropsychological tests (presented in alphabetical order) were used in the study.

Benton Facial Recognition Test (Benton et al. 1983): this thesis used the short version. Subjects are shown a series of faces and asked to pick the identical face from a set of six presented below it, and subsequently are shown a series of faces and asked to pick three faces from a set of six which are the same subject as the target face but with slightly different orientations or lighting. Faces are black and white and distinguishing features such as hair are minimal. The final score on this test is age- and education-adjusted and scores can be classified as normal or abnormal.

Category fluency (Benton and Hamsher 1978): subjects are asked to generate the names of as many animals as possible in one minute.

Digit span (Wechsler 1981): subjects are read out a series of numbers with each series increasing in length, and required to repeat each one back to the experimenter. For the first set of trials the subjects must repeat the numbers verbatim; for the second set subjects are required to reverse the order of the numbers. There are usually two or three trials for each series length, with testing discontinued if the subject fails to repeat all of the numbers at a given length correctly. Alternative versions can be generated by using different numbers. There is no time limit.

Ekman Pictures of Facial Affect (Ekman and Friesen 1976; after Gray et al. 1997): this thesis used the set of 24 pictures of facial affect used by Gray *et al.* in

a similar study, in order to aid comparison between results. This subset was preferred to larger subsets of 60 faces because it required less time to administer. The set consists of black and white pictures of people portraying happiness, sadness, surprise, disgust, anger and fear, with four exemplars of each emotion. Pictures are presented in a random order, with the constraint that three pictures of the same emotion could not appear in sequence, and the names of all six emotions were printed in a pseudo-random order underneath each face. Subjects were asked to pick the emotion which best matched the face. There was no time limit.

Graded Naming Test (GNT) (McKenna and Warrington 1983): a series of 30 line drawings, of increasingly unfamiliar objects, is presented to the subject who is asked to name them. In the event of the subject's making a perceptual error the subject is given the opportunity to answer again. The task is not timed.

Homophone Meaning Generation Test (Warrington 2000): subjects are asked to generate as many homophones as possible for a set of four words, e.g. "bear", "tip". There are two alternative versions and the task is not timed.

Hopkins Verbal Learning Test (Brandt and Benedict 2001): subjects are read a list of 12 words, which fit into three semantic categories, and then asked to recall as many words as possible. This is repeated three times to give an immediate recall score. Twenty–twenty-five minutes later subjects are again asked to recall the words (generating a delayed score and percentage recall) and then asked to give yes/no recognition responses to a list of 24 words read out by the examiner, giving a discrimination score (true positives - false positives). There are six

alternative versions, two of which were used in the current study, and this is not timed.

Letter cancellation: this is an in-house timed pencil and paper task consisting of a page of regularly spaced capital letters ("A" to "E"). Subjects are asked to cross out all of the "A"s as fast as they can.

National Adult Reading Test (NART) (Nelson and Willison 1991): this task requires subjects to read a list of irregular English words ranging from e.g. "CHORD" to "DEMESNE", and in diseases in which reading aloud is relatively spared, score tends to correlate with pre-morbid IQ. There is no time limit.

Phonemic (letter) fluency (Benton and Hamsher 1978): subjects have three one-minute periods in which to generate as many words beginning with "F", "A" and "S" as they can.

Recognition Memory Test (RMT) (Warrington 1984): subjects are shown 50 faces (or words) sequentially and asked to make a yes/no judgement as to whether the face (or word) is pleasant or unpleasant. Subsequently subjects are presented with a two alternative forced choice test in which they have to pick out the target face (or word) from a pair, one of which was not in the original stimulus set. There are two alternative versions. There is no time limit.

Spot the Word Test (Baddeley et al. 1993): subjects are shown 60 word – non-word pairs and asked to circle the real word. Real words tend to be uncommon and score correlates highly with IQ. There are two alternative versions. There is no time limit.

Stroop Test (Stroop 1935; Delis et al. 2001): this is a timed test in which subjects first read out colour names printed in incongruous coloured inks, and then have to name the ink colour of incongruous colour-words (this version was used in chapter 3). In later versions subjects are asked to name colour patches, then to read out colour names printed in black ink, and then to name the ink colour of incongruous colour-words (this version was used in the rest of the thesis).

Symbol Digit Modalities Test (SDMT) (Smith 1968): this is a timed pencil and paper task in which subjects follow a key to fill in digits underneath a set of symbols, as fast as they can.

Trail Making (D-KEFS) (Delis et al. 2001): this is a timed pencil and paper task comprising five separate sections for number cancellation, number sequencing, letter sequencing, alternating sequencing, and motor speed.

Trail Making Test for Adults (Reitan and Wolfson 2004): this is a timed pencil and paper task with two parts; in part A subjects draw a line to connect 25 numbers in sequence, and in part B they must draw a line to connect 25 numbers and letters in alternating sequence.

Visual Object and Space Perception Battery (VOSP) (Warrington and James 1991): the silhouette (objects) subtest was used in which subjects are asked to identify a series of 15 black silhouettes of everyday objects, which increase in unfamiliarity.

APPENDIX 4: COGNITIVE SCORES FROM CHAPTER 4

Table A4-1: Mean (SD) neuropsychological performance at baseline for the control and early HD groups, with differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ

	Control	Early HD	Difference (Early HD - Control)	
Test	N=20	N=40	Crude	Adjusted
UHDRS cognitive (n)	334.8 (33.8)	231.5 (53.6)	-103.3 (-125.9, -80.7) p<0.0001	-98.2 (-119.2, -77.2) p<0.0001
Stroop colour (n)	80.25 (9.3)	57.0 (12.7)	-23.2 (-29.0, -17.5) p<0.0001	-22.8 (-28.5, -17.2) p<0.0001
Stroop word (n)	107.3 (14.0)	75.4 (21.2)	-31.9 (-41.0, -22.7) p<0.0001	-29.9 (-38.7, -21.1) p<0.0001
Stroop interference (n)	45.9 (7.5)	32.8 (9.2)	-13.1 (-17.5, -8.7) p<0.0001	-12.0 (-16.1, -7.8) p<0.0001
SDMT (n)	55.6 (7.8)	33.3 (10.3)	-22.3 (-27.0, -17.5) p<0.0001	-21.0 (-25.6, -16.4) p<0.0001
Phonemic fluency (n)	45.9 (9.0)	33.0 (11.3)	-12.9 (-18.2, -7.5) p<0.0001	-12.5 (-17.4, -7.6) p<0.0001
Category fluency (n)	24.1 (3.9)	17.9 (5.6)	-6.2 (-8.7, -3.7) p<0.0001	-5.8 (-8.2, -3.3) p<0.0001
HMGT (n)	11.9 (2.2)	10.3 (2.5)	-1.6 (-2.8, -0.4) p=0.01	-1.4 (-2.5, -0.4) p=0.007
TMT A (sec)	21.3 (6.5)	37.6 (11.5)	-16.2 (-20.6, -11.7) p<0.05	-14.8 (-19.5, -10.4) p<0.05

Test	Control N=20	Early HD N=40	Difference (Early HD - Control) Crude	Adjusted
TMT B (sec)	62.4 (41.0)	119.6 (55.6)	-57.2 (-82.2, -30.7) p<0.05	-53.7 (-79.5, -29.3) p<0.05
TMT B-A (sec)	41.1 (40.1)	83.6 (51.5)	-42.5 (-64.1, -19.1) p<0.05	-39.8 (-65.1, -15.8) p<0.05
TMT B/A	3.0 (1.9)	3.4 (1.3)	-0.4 (-1.2, 0.7) p>0.05	-0.3 (-1.2, 0.8) p>0.05
Cancelling As (sec)	17.3 (4.2)	28.1 (8.5)	-10.8 (-14.1, -7.5) p<0.0001	-10.2 (-13.5, -7.0) p<0.0001
RT near (msec)	654 (83)	932 (175)	-277.4 (-342.9, -211.1) p<0.05	-260.0 (-322.4, -195.4) p<0.05
RT far (msec)	711 (81)	980 (178)	-269.0 (-333.0, -203.7) p<0.05	-253.8 (-321.2, -191.1) p<0.05
RT return from near (msec)	290 (95)	594 (248)	-303.9 (-390.3, -215.1) p<0.05	-274.6 (-355.5, -190.7) p<0.05
RT return from far (msec)	299 (112)	593 (233)	-294.4 (-388.9, -212.1) p<0.05	-271.1 (-361.5, -180.8) p<0.05
HVLT immediate (/36)	25.6 (6.1)	19.7 (5.2)	-5.9 (-9.0, -2.7) p<0.0001	-5.1 (-8.1, -2.1) p=0.001
HVLT delayed (/12)	9.7 (2.2)	6.4 (2.4)	-3.2 (-4.4, -2.0) p<0.0001	-2.9 (-4.0, -1.7) p<0.0001
HVLT % recalled	95.7 (14.5)	80.2 (24.0)	-15.5 (-25.4, -5.6) p=0.003	-13.1 (-23.0, -3.2) p=0.01
HVLT discrimination (/12)	11.1 (1.2)	8.8 (2.3)	-2.3 (-3.2, -1.4) p<0.0001	-2.1 (-2.9, -1.2) p<0.0001

Test	Control N=20	Early HD N=40	Difference (Early HD - Control) Crude	Adjusted
Digit span forwards (/21)	13.4 (3.8)	10.5 (3.2)	-3.0 (-4.9, -1.0) p=0.004	-2.7 (-4.3, -1.0) p=0.002
Digit span backwards (/21)	12.5 (4.1)	8.7 (3.8)	-3.8 (-6.0, -1.6) p=0.001	-3.5 (-5.5, -1.6) p<0.0001
RMF (/50)	42.7 (4.0)	34.8 (7.0)	-7.9 (-10.8, -5.1) p<0.0001	-7.5 (-10.5, -4.6) p<0.0001
RMW (/50)	48.2 (2.5)	43.4 (4.7)	-4.8 (-6.5, -2.9) p<0.05	-4.3 (-5.5, -2.7) p<0.05
RMT difference	-5.5 (4.6)	-8.6 (6.1)	-3.1 (-5.9, -0.3) p=0.03	-3.3 (-6.2, -0.3) p=0.029
GNT (/30)	21.4 (3.7)	20.7 (4.7)	-0.7 (-2.9, 1.5) p=0.54	-0.7 (-2.2, 0.7) p=0.32
Naming latency (ln(sec))	0.6 (0.4)	0.8 (0.4)	-0.3 (-0.4, -0.2) p<0.0001	-0.2 (-0.3, -0.1) p<0.0001
VOSP (/15)	8.4 (2.6)	7.5 (2.7)	-0.8 (-2.1, 0.4) p=0.20	-0.6 (-1.8, 0.6) p=0.20
Benton (/54)	48.0 (3.8)	43.5 (4.7)	-4.5 (-6.8, -2.2) p<0.0001	
Emotion recognition (/24)	21.4 (1.5)	16.8 (3.3)	-4.5 (-5.7, -3.2) p<0.05	-4.3 (-23.6, -12.9) p<0.05

Negative differences indicate a lower score or slower speed (longer time taken in timed tasks) in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models; the Benton scores are already age- and education-adjusted; where 95% bootstrapped CIs were used precise p values were not obtained.

Table A4-2: Mean (SD) neuropsychological performance at baseline for the control and premanifest groups, with differences (95% confidence intervals) with and without adjustment for age and estimated premorbid IQ

	Control	PM	Difference (PM - Control)	
Test	N=20	N=21	Crude	Adjusted
UHDRS cognitive (n)	334.8 (33.8)	321.9 (51.1)	-12.9 (-39.6, 13.8) p=0.34	-14.4 (-37.4, 8.6) p=0.22
Stroop colour (n)	80.25 (9.3)	77.5 (10.5)	-2.8 (-8.9, 3.4) p=0.37	-2.9 (-9.1, 3.3) p=0.36
Stroop word (n)	107.3 (14.0)	104.1 (19.8)	-3.2 (-13.7, 7.4) p=0.55	-3.8 (-13.1, 5.5) p=0.42
Stroop interference (n)	45.9 (7.5)	47.8 (9.5)	1.9 (-3.4, 7.2) p=0.47	1.1 (-4.2, 6.3) p=0.69
SDMT (n)	55.6 (7.8)	53.4 (10.3)	-2.1 (-7.8, 3.5) p=0.46	-3.3 (-8.8, 2.2) p=0.24
Letter fluency (n)	45.9 (9.0)	39.1 (13.5)	-6.8 (-13.8, 0.3) p=0.06	-5.5 (-12.5, 1.6) p=0.13
Category fluency (n)	24.1 (3.9)	23.0 (7.7)	-1.0 (-4.7, 2.7) p=0.59	-0.8 (-4.1, 2.5) p=0.64
HMGT (n)	11.9 (2.2)	11.6 (2.7)	-0.3 (-1.7, 1.1) p=0.67	0.0 (-1.1, 1.1) p=1.0
TMT A (sec)	21.3 (6.5)	22.4 (5.0)	-1.1 (-4.9, 2.5) p>0.05	-3.3 (-1.1, 7.9) p>0.05
TMT B (sec)	62.4 (41.0)	60.2 (25.1)	2.3 (-14.8, 26.5) p>0.05	-2.5 (-19.7, 17.3) p>0.05

Test	Control N=20	PM N=21	Difference (PM - Control) Crude	Adjusted
TMT B-A (sec)	41.1 (40.1)	37.7 (22.3)	3.4 (-13.8, 26.7) p>0.05	0.1 (-14.5, 18.4) p>0.05
TMT B/A	3.0 (1.9)	2.7 (0.8)	0.4 (-0.4, 1.4) p>0.05	0.3 (-0.3, 1.4) p>0.05
Cancelling As (sec)	17.3 (4.2)	18.5 (4.5)	1.2 (-1.5, 3.9) p=0.37	-2.0 (-5.0, 1.0) p=0.18
RT near (msec)	654 (83)	641 (91)	12.7 (-38.2, 66.3) p>0.05	-7.7 (62.1, 39.7) p>0.05
RT far (msec)	711 (81)	690 (96)	21.7 (-35.9, 75.4) p>0.05	-0.8 (-58.3, 51.9) p>0.05
RT return from near (msec)	290 (95)	357 (138)	-67.9 (-136.5, 5.0) p>0.05	-118.4 (-206.3, -45.2) p<0.05
RT return from far (msec)	299 (112)	356 (143)	-57.2 (-129.3, 23.6) p>0.05	-99.3 (-183.6, -26.5) p<0.05
HVLT immediate (/36)	25.6 (6.1)	27.0 (4.9)	1.5 (-2.0, 4.9) p=0.40	0.7 (-2.3, 3.6) p=0.66
HVLT delayed (/12)	9.7 (2.2)	9.2 (2.2)	-0.4 (-1.8, 1.0) p=0.55	-0.8 (-2.2, 0.6) p=0.24
HVLT % recalled	95.7 (14.5)	89.5 (15.6)	-6.2 (-15.5, 3.1) p=0.19	-9.1 (-20.0, 1.7) p=0.10
HVLT discrimination (/12)	11.1 (1.2)	10.7 (1.2)	-0.4 (-1.1, 0.3) p=0.28	-0.7 (-1.4, 0.1) p=0.10
Digit span forwards (/21)	13.4 (3.8)	13.3 (3.2)	-0.1 (-2.2, 2.1) p=0.95	0.02 (-2.0, 2.1) p=0.98

	Control	PM	Difference (PM - Control)	
Test	N=20	N=21	Crude	Adjusted
Digit span backwards (/21)	12.5 (4.1)	11.5 (4.0)	-1.0 (-3.5, 1.5) p=0.44	-0.9 (-3.2, 1.4) p=0.43
RMF (/50)	42.7 (4.0)	41.3 (5.3)	-1.4 (-4.3, 1.5) p=0.35	-1.5 (-4.6, 1.7) p=0.36
RMW (/50)	48.2 (2.5)	48.2 (1.9)	0.04 (-1.3, 1.4) p>0.05	-0.3 (-1.4, 0.9) p<0.05
RMT difference	-5.5 (4.6)	-6.9 (5.5)	-1.4 (-4.5, 1.7) p=0.38	-1.2 (-4.5, 2.1) p=0.48
GNT (/30)	21.4 (3.7)	21.8 (4.5)	0.4 (-2.2, 2.9) p=0.78	1.7 (0.01, 3.3) p=0.049
Naming latency (ln(msec)	0.6 (0.4)	0.7 (0.4)	-0.1 (-0.2, 0.05) p>0.05	-0.1 (-0.2, 0.03) p>0.05
VOSP (/15)	8.4 (2.6)	7.5 (3.5)	-0.8 (-2.6, 0.9) p=0.34	-1.1 (-2.7, 0.6) p=0.20
Benton (/54)	48.0 (3.8)	47.7 (3.7)	-0.3 (-2.6, 2.1) p=0.81	
Emotion recognition (/24)	21.4 (1.5)	20.6 (2.3)	-0.7 (-2.0, 0.5) p>0.05	-2.7 (-7.9, 3.0) p>0.05

Negative differences indicate a lower score or slower speed (longer time taken in timed tasks) in HD subjects relative to controls; crude differences are the absolute difference between groups before adjustment for age and IQ; adjusted differences are the differences having adjusted for age and IQ by including them as covariates in linear regression models; the Benton scores are already age- and education-adjusted; where 95% bootstrapped CIs were used precise p values were not obtained.

REFERENCES

Acer, N., Sahin, B., Bas, O., Ertekin, T., Usanmaz, M., 2007. Comparison of three methods for the estimation of total intracranial volume: stereologic, planimetric, and anthropometric approaches. Ann Plast Surg 58: 48-53.

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., Damasio, A.R., 2005. A mechanism for impaired fear recognition after amygdala damage. Nature 433: 68-72.

Alexander, G.E., DeLong, M.R., Strick, P.L., 1986. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9: 357-381.

Andrews, T.C., Weeks, R.A., Turjanski, N. *et al.*, 1999. Huntington's disease progression. PET and clinical observations. Brain 122 (Pt 12): 2353-2363.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38: 95-113.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry--the methods. Neuroimage 11: 805-821.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26: 839-851.

Aylward, E.H., 2007. Change in MRI striatal volumes as a biomarker in preclinical Huntington's disease. Brain Res Bull 72: 152-158.

Aylward, E.H., Anderson, N.B., Bylsma, F.W. *et al.*, 1998. Frontal lobe volume in patients with Huntington's disease. Neurol 50: 252-258.

Aylward, E.H., Brandt, J., Codori, A.M., Mangus, R.S., Barta, P.E., Harris, G.J., 1994. Reduced basal ganglia volume associated with the gene for Huntington's disease in asymptomatic at-risk persons. Neurol 44: 823-828.

Aylward, E.H., Codori, A.M., Barta, P.E., Pearlson, G.D., Harris, G.J., Brandt, J., 1996. Basal ganglia volume and proximity to onset in presymptomatic Huntington disease. Arch Neurol 53: 1293-1296.

Aylward, E.H., Codori, A.M., Rosenblatt, A. *et al.*, 2000. Rate of caudate atrophy in presymptomatic and symptomatic stages of Huntington's disease. Mov Disord 15: 552-560.

Aylward, E.H., Li, Q., Stine, O.C. *et al.*, 1997. Longitudinal change in basal ganglia volume in patients with Huntington's disease. Neurol 48: 394-399.

Aylward, E.H., Rosenblatt, A., Field, K. *et al.*, 2003. Caudate volume as an outcome measure in clinical trials for Huntington's disease: a pilot study. Brain Res Bull 62: 137-141.

Aylward, E.H., Sparks, B.F., Field, K.M. *et al.*, 2004. Onset and rate of striatal atrophy in preclinical Huntington disease. Neurol 63: 66-72.

Bachoud-Lévi, A.C., Maison, P., Bartolomeo, P. *et al.*, 2001. Retest effects and cognitive decline in longitudinal follow-up of patients with early HD. Neurol 56: 1052-1058.

Bäckman, L., Robins-Wahlin, T.B., Lundin, A., Ginovart, N., Farde, L., 1997. Cognitive deficits in Huntington's disease are predicted by dopaminergic PET markers and brain volumes. Brain 120 (Pt 12): 2207-2217.

Baddeley, A., Emslie, H., Nimmo-Smith, I., 1993. The Spot-the-Word test: a robust estimate of verbal intelligence based on lexical decision. Br J Clin Psychol 32 (Pt 1): 55-65.

Balleine, B.W., Delgado, M.R., Hikosaka, O., 2007. The role of the dorsal striatum in reward and decision-making. J Neurosci 27: 8161-8165.

Bamford, K.A., Caine, E.D., Kido, D.K., Cox, C., Shoulson, I., 1995. A prospective evaluation of cognitive decline in early Huntington's disease: functional and radiographic correlates. Neurol 45: 1867-1873.

Bamford, K.A., Caine, E.D., Kido, D.K., Plassche, W.M., Shoulson, I., 1989. Clinical-pathologic correlation in Huntington's disease: a neuropsychological and computed tomography study. Neurol 39: 796-801.

Banich, M.T., Milham, M.P., Atchley, R. *et al.*, 2000. fMri studies of Stroop tasks reveal unique roles of anterior and posterior brain systems in attentional selection. J Cogn Neurosci 12: 988-1000.

Barrios, F.A., Gonzalez, L., Favila, R. *et al.*, 2007. Olfaction and neurodegeneration in HD. Neuroreport 18: 73-76.

Bates, G.P., 2005. History of genetic disease: the molecular genetics of Huntington disease - a history. Nat Rev Genet 6: 766-773.

Beck AT, Steer RA, 1993. Manual for the Beck Depression Inventory. San Antonio, TX: Psychological Corporation.

Beglinger, L.J., Nopoulos, P.C., Jorge, R.E. *et al.*, 2005. White Matter Volume and Cognitive Dysfunction in Early Huntington's Disease. Cogn Behav Neurol 18: 102-107.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57: 289-300.

Benton AL, Hamsher Kd, 1978. Multilingual Aphasia Examination. Iowa City: University of Iowa.

Benton AL, Hamsher KS, Varney N, Spreen O, 1983. Contributions to neuropsychological assessment: A clinical manual. Oxford: Oxford University Press.

Berrios, G.E., Wagle, A.C., Markova, I.S. *et al.*, 2001. Psychiatric symptoms and CAG repeats in neurologically asymptomatic Huntington's disease gene carriers. Psychiatry Res 102: 217-225.

Biomarkers Definitions Working Group, 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69: 89-95.

Blair, R.J., Morris, J.S., Frith, C.D., Perrett, D.I., Dolan, R.J., 1999. Dissociable neural responses to facial expressions of sadness and anger. Brain 122 (Pt 5): 883-893.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1: 307-310.

Blekher, T., Johnson, S.A., Marshall, J. *et al.*, 2006. Saccades in presymptomatic and early stages of Huntington disease. Neurol 67: 394-399.

Bonelli, R.M., Cummings, J.L., 2007. Frontal-subcortical circuitry and behavior. Dialogues Clin Neurosci 9: 141-151.

Brandt J. Cognitive impairments in Huntington's disease: insights into the neuropsychology of the striatum. In: Boller F, Grafman J, editors. Handbook of Neuropsychology. New York: Elsevier; 1993. p. 241-64.

Brandt J, Benedict RHB, 2001. Hopkins Verbal Learning Test - Revised. Professional Manual. Lutz, FL: Psychological Assessment Resources, Inc..

Brandt, J., Bylsma, F.W., Aylward, E.H., Rothlind, J., Gow, C.A., 1995. Impaired source memory in Huntington's disease and its relation to basal ganglia atrophy. J Clin Exp Neuropsychol 17: 868-877.

Brandt, J., Bylsma, F.W., Gross, R., Stine, O.C., Ranen, N., Ross, C.A., 1996. Trinucleotide repeat length and clinical progression in Huntington's disease. Neurol 46: 527-531.

Buchel, C., Morris, J., Dolan, R.J., Friston, K.J., 1998. Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron 20: 947-957.

Calder, A.J., Keane, J., Lawrence, A.D., Manes, F., 2004. Impaired recognition of anger following damage to the ventral striatum. Brain 127: 1958-1969.

Calder, A.J., Keane, J., Manes, F., Antoun, N., Young, A.W., 2000. Impaired recognition and experience of disgust following brain injury. Nat Neurosci 3: 1077-1078.

Calder, A.J., Keane, J., Manly, T. *et al.*, 2003. Facial expression recognition across the adult life span. Neuropsychologia 41: 195-202.

Calder, A.J., Lawrence, A.D., Young, A.W., 2001. Neuropsychology of fear and loathing. Nat Rev Neurosci 2: 352-363.

Calder, A.J., Young, A.W., Rowland, D., Perrett, D.I., Hodges, J.R., Etcoff, N.L., 1996. Facial emotion recognition after bilateral amygdala damage: Differentially severe impairment of fear. Cognitive Neuropsychology 13: 699-745.

Cattaneo, E., Zuccato, C., Tartari, M., 2005. Normal huntingtin function: an alternative approach to Huntington's disease. Nat Rev Neurosci 6: 919-930.

Ciarmiello, A., Cannella, M., Lastoria, S. *et al.*, 2006. Brain white-matter volume loss and glucose hypometabolism precede the clinical symptoms of Huntington's disease. J Nucl Med 47: 215-222.

Costa, R.M., 2007. Plastic corticostriatal circuits for action learning: what's dopamine got to do with it? Ann N Y Acad Sci 1104: 172-191.

Craufurd D, Snowden J. Neuropsychological and neuropsychiatric aspects of Huntington's disease. In: Bates G, Harper P, Jones L, editors. Huntington's Disease. Oxford: Oxford University Press; 2002. p. 62-94.

Craufurd, D., Thompson, J.C., Snowden, J.S., 2001. Behavioral changes in Huntington Disease. Neuropsychiatry Neuropsychol Behav Neurol 14: 219-226.

Cummings, J.L., 1993. Frontal-subcortical circuits and human behavior. Arch Neurol 50: 873-880.

de Boo, G.M., Tibben, A., Lanser, J.B. *et al.*, 1997. Early cognitive and motor symptoms in identified carriers of the gene for Huntington disease. Arch Neurol 54: 1353-1357.

de Gelder, B., Van den, S.J., de Diego, B.R., Bachoud-Levi, A.C., 2008. Huntington's disease impairs recognition of angry and instrumental body language. Neuropsychologia 46: 369-373.

de la Monte, S.M., Vonsattel, J.P., Richardson, E.P., Jr., 1988. Morphometric demonstration of atrophic changes in the cerebral cortex, white matter, and neostriatum in Huntington's disease. J Neuropathol Exp Neurol 47: 516-525.

Delgado, M.R., 2007. Reward-related responses in the human striatum. Ann N Y Acad Sci 1104: 70-88.

Delis DC, Kaplan EF, Kramer JH, 2001. Delis-Kaplan Executive Function System. San Antonia, Texas, USA: The Psychological Corporation.

Douaud, G., Gaura, V., Ribeiro, M.J. *et al.*, 2006. Distribution of grey matter atrophy in Huntington's disease patients: a combined ROI-based and voxel-based morphometric study. Neuroimage 32: 1562-1575.

Duning, T., Kloska, S., Steinstrater, O., Kugel, H., Heindel, W.. Knecht, S., 2005. Dehydration confounds the assessment of brain atrophy. Neurol 64: 548-550.

Duyao, M., Ambrose, C., Myers, R. *et al.*, 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet 4: 387-392.

Ekman P, Friesen WV, 1976. Pictures of facial affect. Palo Alto, CA: Consulting Psychologists Press.

Feigin, A., Ghilardi, M.F., Huang, C. *et al.*, 2006. Preclinical Huntington's disease: Compensatory brain responses during learning. Ann Neurol 59: 53-59.

Feigin, A., Tang, C., Ma, Y. *et al.*, 2007. Thalamic metabolism and symptom onset in preclinical Huntington's disease. Brain 130: 2858-2867.

Fennema-Notestine, C., Archibald, S.L., Jacobson, M.W. *et al.*, 2004. In vivo evidence of cerebellar atrophy and cerebral white matter loss in Huntington disease. Neurol 63: 989-995.

Feudjo-Tepie, Frost C, Wang D, Bakhai A. Missing Data. In: Wang D, Bakhai A, editors. Clinical trials: a practical guide to design, analysis, and reporting. London: Remedica; 2006. p. 339-52.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12: 189-198.

Fox, N.C., Black, R.S., Gilman, S. *et al.*, 2005. Effects of Abeta immunization (AN1792) on MRI measures of cerebral volume in Alzheimer disease. Neurol 64: 1563-1572.

Fox, N.C., Cousens, S., Scahill, R., Harvey, R.J., Rossor, M.N., 2000. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects [see comments]. Arch Neurol 57: 339-344.

Fox, N.C., Freeborough, P.A., 1997. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. J Magn Reson Imaging 7: 1069-1075.

Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans Med Imaging 16: 623-629.

Freeborough, P.A., Fox, N.C., Kitney, R.I., 1997. Interactive algorithms for the segmentation and quantitation of 3-D MRI brain scans. Computer Methods and Programs in Biomedicine 53: 15-25.

Freeborough, P.A., Woods, R.P., Fox, N.C., 1996. Accurate registration of serial 3D MR brain images and its application to visualizing change in neurodegenerative disorders. J Comput Assist Tomogr 20: 1012-1022.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S., 1995. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Hum Brain Mapp 2: 189-210.

Furtado, S., Sossi, V., Hauser, R.A. *et al.*, 2005. Positron emission tomography after fetal transplantation in Huntington's disease. Ann Neurol 58: 331-337.

Furtado, S., Suchowersky, O., Rewcastle, B., Graham, L., Klimek, M.L., Garber, A., 1996. Relationship between trinucleotide repeats and neuropathological changes in Huntington's disease. Ann Neurol 39: 132-136.

Gaura, V., Bachoud-Levi, A.C., Ribeiro, M.J. *et al.*, 2004. Striatal neural grafting improves cortical metabolism in Huntington's disease patients. Brain 127: 65-72.

Gavazzi, C., Nave, R.D., Petralli, R. *et al.*, 2007. Combining functional and structural brain magnetic resonance imaging in Huntington disease. J Comput Assist Tomogr 31: 574-580.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15: 870-878.

Georgiou-Karistianis, N., Sritharan, A., Farrow, M. *et al.*, 2007. Increased cortical recruitment in Huntington's disease using a Simon task. Neuropsychologia 45: 1791-1800.

Ghilardi, M.F., Silvestri, G., Feigin, A. *et al.*, 2008. Implicit and explicit aspects of sequence learning in pre-symptomatic Huntington's disease. Parkinsonism Relat Disord.

Giordani, B., Berent, S., Boivin, M.J. *et al.*, 1995. Longitudinal neuropsychological and genetic linkage analysis of persons at risk for Huntington's disease. Arch Neurol 52: 59-64.

Golding, C.V., Danchaivijitr, C., Hodgson, T.L., Tabrizi, S.J., Kennard, C., 2006. Identification of an oculomotor biomarker of preclinical Huntington disease. Neurol 67: 485-487.

Good, C.D., Johnsrude, I., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001a. Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. Neuroimage 14: 685-700.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001b. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 14: 21-36.

Gorno-Tempini, M.L., Pradelli, S., Serafini, M. *et al.*, 2001. Explicit and incidental facial expression processing: an fMRI study. Neuroimage 14: 465-473.

Gray, J.M., Young, A.W., Barker, W.A., Curtis, A., Gibson, D., 1997. Impaired recognition of disgust in Huntington's disease gene carriers. Brain 120: 2029-2038.

Gutenkst C, Norflus F, Hersch SM. The neuropathology of Huntington's disease. In: Bates G, Harper P, Jones L, editors. Huntington's Disease. Oxford: Oxford University Press; 2002. p. 251-75.

Haber, S.N., 2003. The primate basal ganglia: parallel and integrative networks. J Chem Neuroanat 26: 317-330.

Halliday, G.M., McRitchie, D.A., Macdonald, V., Double, K.L., Trent, R.J., McCusker, E., 1998. Regional specificity of brain atrophy in Huntington's disease. Exp Neurol 154: 663-672.

Handley, O.J., Naji, J.J., Dunnett, S.B., Rosser, A.E., 2006. Pharmaceutical, cellular and genetic therapies for Huntington's disease. Clin Sci (Lond) 110: 73-88.

Hanes, K.R., 1996. Brief Report: Bradyphrenia in Parkinson's Disease, Huntington's Disease, and Schizophrenia. Cognit Neuropsychiatry 1: 165-170.

Harper P. The epidemiology of Huntington's disease. In: Bates G, Harper P, Jones L, editors. Huntington's Disease. Oxford: Oxford University Press; 2002. p. 3-27.

Harris, G.J., Aylward, E.H., Peyser, C.E. *et al.*, 1996. Single photon emission computed tomographic blood flow and magnetic resonance volume imaging of basal ganglia in Huntington's disease. Arch Neurol 53: 316-324.

Harris, G.J., Codori, A.M., Lewis, R.F., Schmidt, E., Bedi, A., Brandt, J., 1999. Reduced basal ganglia blood flow and volume in pre-symptomatic, gene-tested persons at-risk for Huntington's disease. Brain 122 (Pt 9): 1667-1678.

Harris, G.J., Pearlson, G.D., Peyser, C.E. *et al.*, 1992. Putamen volume reduction on magnetic resonance imaging exceeds caudate changes in mild Huntington's disease. Ann Neurol 31: 69-75.

Harrison, B.J., Shaw, M., Yucel, M. *et al.*, 2005. Functional connectivity during Stroop task performance. Neuroimage 24: 181-191.

Hayes, C.J., Stevenson, R.J., Coltheart, M., 2007. Disgust and Huntington's disease. Neuropsychologia 45: 1135-1151.

Henderson, J.M., Williams, C.C., Falk, R.J., 2005. Eye movements are functional during face learning. Mem Cognit 33: 98-106.

Hennenlotter, A., Schroeder, U., Erhard, P. *et al.*, 2004. Neural correlates associated with impaired disgust processing in pre-symptomatic Huntington's disease. Brain 127: 1446-1453.

Henry, J.D., Crawford, J.R., Phillips, L.H., 2005. A meta-analytic review of verbal fluency deficits in Huntington's disease. Neuropsychology 19: 243-252.

Henson, R.N., Rugg, M.D., Shallice, T., Dolan, R.J., 2000. Confidence in recognition memory for words: dissociating right prefrontal roles in episodic retrieval. J Cogn Neurosci 12: 913-923.

Hirshorn, E.A., Thompson-Schill, S.L., 2006. Role of the left inferior frontal gyrus in covert word retrieval: neural correlates of switching during verbal fluency. Neuropsychologia 44: 2547-2557.

Ho, A.K., Nestor, P.J., Williams, G.B. *et al.*, 2004. Pseudo-neglect in Huntington's disease correlates with decreased angular gyrus density. Neuroreport 15: 1061-1064.

Ho, A.K., Sahakian, B.J., Brown, R.G. *et al.*, 2003. Profile of cognitive progression in early Huntington's disease. Neurol 61: 1702-1706.

Ho, A.K., Sahakian, B.J., Robbins, T.W., Barker, R.A., Rosser, A.E., Hodges, J.R., 2002. Verbal fluency in Huntington's disease: a longitudinal analysis of phonemic and semantic clustering and switching. Neuropsychologia 40: 1277-1284.

Hodges, J.R., Salmon, D.P., Butters, N., 1991. The nature of the naming deficit in Alzheimer's and Huntington's disease. Brain 114 (Pt 4): 1547-1558.

Hogarth, P., Kayson, E., Kieburtz, K. *et al.*, 2005. Interrater agreement in the assessment of motor manifestations of Huntington's disease. Mov Disord 20: 293-297.

Huntington Study Group, 1996. Unified Huntington's Disease Rating Scale: reliability and consistency. Huntington Study Group. Mov Disord 11: 136-142.

Huntington's Disease Collaborative Research Group, 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72: 971-983.

Huntington, G., 1872. On Chorea. The Medical and Surgical Reporter: A Weekly Journal 26: 317-321.

Jack, C.R., Jr., Shiung, M.M., Gunter, J.L. *et al.*, 2004. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurol 62: 591-600.

Jacobs, D.H., Shuren, J., Heilman, K.M., 1995. Impaired perception of facial identity and facial affect in Huntington's disease. Neurol 45: 1217-1218.

Jason, G.W., Suchowersky, O., Pajurkova, E.M. *et al.*, 1997. Cognitive manifestations of Huntington disease in relation to genetic structure and clinical onset. Arch Neurol 54: 1081-1088.

Jech, R., Klempir, J., Vymazal, J. *et al.*, 2007. Variation of selective gray and white matter atrophy in Huntington's disease. Mov Disord 22: 1783-1789.

Joel, D., 2001. Open interconnected model of basal ganglia-thalamocortical circuitry and its relevance to the clinical syndrome of Huntington's disease. Mov Disord 16: 407-423.

Johnson, S.A., Stout, J.C., Solomon, A.C. *et al.*, 2007. Beyond disgust: impaired recognition of negative emotions prior to diagnosis in Huntington's disease. Brain 130: 1732-1744.

Josiassen, R.C., Curry, L.M., Mancall, E.L., 1983. Development of neuropsychological deficits in Huntington's disease. Arch Neurol 40: 791-796.

Julien, C.L., Thompson, J.C., Wild, S. *et al.*, 2007. Psychiatric disorders in preclinical Huntington's disease. J Neurol Neurosurg Psychiatry 78: 939-943.

Karas, G.B., Burton, E.J., Rombouts, S.A. *et al.*, 2003. A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18: 895-907.

Kassubek, J., Bernhard, L.G., Ecker, D. *et al.*, 2004a. Global cerebral atrophy in early stages of Huntington's disease: quantitative MRI study. Neuroreport 15: 363-365.

Kassubek, J., Gaus, W., Landwehrmeyer, G.B., 2004b. Evidence for more widespread cerebral pathology in early HD: an MRI-based morphometric analysis. Neurol 62: 523-524.

Kassubek, J., Juengling, F.D., Ecker, D., Landwehrmeyer, G.B., 2005. Thalamic atrophy in Huntington's disease co-varies with cognitive performance: a morphometric MRI analysis. Cereb Cortex 15: 846-853.

Kassubek, J., Juengling, F.D., Kioschies, T. *et al.*, 2004c. Topography of cerebral atrophy in early Huntington's disease: a voxel based morphometric MRI study. J Neurol Neurosurg Psychiatry 75: 213-220.

Keller, S.S., Wilke, M., Wieshmann, U.C., Sluming, V.A., Roberts, N., 2004. Comparison of standard and optimized voxel-based morphometry for analysis of brain changes associated with temporal lobe epilepsy. Neuroimage 23: 860-868.

Kieburtz K, Shoulson I. Therapeutic trials in Huntington's disease. In: Bates G, Harper P, Jones L, editors. Huntington's Disease. Oxford: Oxford University Press; 2002. p. 501-11.

Kim, J.S., Reading, S.A., Brashers-Krug, T., Calhoun, V.D., Ross, C.A., Pearlson, G.D., 2004. Functional MRI study of a serial reaction time task in Huntington's disease. Psychiatry Res 131: 23-30.

Kipps, C.M., Duggins, A.J., Mahant, N., Gomes, L., Ashburner, J., McCusker, E.A., 2005. Progression of structural neuropathology in preclinical Huntington's disease: a tensor based morphometry study. J Neurol Neurosurg Psychiatry 76: 650-655.

Kipps, C.M., Duggins, A.J., McCusker, E.A., Calder, A.J., 2007. Disgust and happiness recognition correlate with anteroventral insula and amygdala volume respectively in preclinical Huntington's disease. J Cogn Neurosci 19: 1206-1217.

Kirkwood, S.C., Siemers, E., Hodes, M.E., Conneally, P.M., Christian, J.C., Foroud, T., 2000. Subtle changes among presymptomatic carriers of the Huntington's disease gene. J Neurol Neurosurg Psychiatry 69: 773-779.

Klöppel, S., Draganski, B., Golding, C.V. *et al.*, 2008. White matter connections reflect changes in voluntary-guided saccades in pre-symptomatic Huntington's disease. Brain 131: 196-204.

Krolak-Salmon, P., Henaff, M.A., Isnard, J. *et al.*, 2003. An attention modulated response to disgust in human ventral anterior insula. Ann Neurol 53: 446-453.

Lang, C.J., Majer, M., Balan, P., Reischies, F.M., 2000a. Recall and recognition in Huntington's disease. Arch Clin Neuropsychol 15: 361-371.

Lang, P.J., Davis, M., Ohman, A., 2000b. Fear and anxiety: animal models and human cognitive psychophysiology. J Affect Disord 61: 137-159.

Langbehn, D.R., Brinkman, R.R., Falush, D., Paulsen, J.S., Hayden, M.R., 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin Genet 65: 267-277.

Lawrence, A.D., Hodges, J.R., Rosser, A.E. *et al.*, 1998a. Evidence for specific cognitive deficits in preclinical Huntington's disease. Brain 121: 1329-1341.

Lawrence, A.D., Sahakian, B.J., Hodges, J.R., Rosser, A.E., Lange, K.W., Robbins, T.W., 1996. Executive and mnemonic functions in early Huntington's disease. Brain 119 (Pt 5): 1633-1645.

Lawrence, A.D., Sahakian, B.J., Robbins, T.W., 1998b. Cognitive functions and corticostriatal circuits: insights from Huntington's disease. Trends in Cognitive Sciences 2: 379-389.

Lawrence, A.D., Watkins, L.H., Sahakian, B.J., Hodges, J.R., Robbins, T.W., 2000. Visual object and visuospatial cognition in Huntington's disease: implications for information processing in corticostriatal circuits. Brain 123 (Pt 7): 1349-1364.

Lawrence, A.D., Weeks, R.A., Brooks, D.J. *et al.*, 1998c. The relationship between striatal dopamine receptor binding and cognitive performance in Huntington's disease. Brain 121 (Pt 7): 1343-1355.

Lemiere, J., Decruyenaere, M., Evers-Kiebooms, G., Vandenbussche, E., Dom, R., 2002. Longitudinal study evaluating neuropsychological changes in so-called asymptomatic carriers of the Huntington's disease mutation after 1 year. Acta Neurol Scand 106: 131-141.

Lemiere, J., Decruyenaere, M., Evers-Kiebooms, G., Vandenbussche, E., Dom, R., 2004. Cognitive changes in patients with Huntington's disease (HD) and asymptomatic carriers of the HD mutation--a longitudinal follow-up study. J Neurol 251: 935-942.

Lewis, E.B., Fox, N.C., 2004. Correction of differential intensity inhomogeneity in longitudinal MR images. Neuroimage 23: 75-83.

Ludlow, C.L., Connor, N.P., Bassich, C.J., 1987. Speech timing in Parkinson's and Huntington's disease. Brain Lang 32: 195-214.

Mahant, N., McCusker, E.A., Byth, K., Graham, S., 2003. Huntington's disease: clinical correlates of disability and progression. Neurol 61: 1085-1092.

Mann, D.M., Oliver, R., Snowden, J.S., 1993. The topographic distribution of brain atrophy in Huntington's disease and progressive supranuclear palsy. Acta Neuropathol 85: 553-559.

Marder, K., Zhao, H., Myers, R.H. *et al.*, 2000. Rate of functional decline in Huntington's disease. Huntington Study Group. Neurol 54: 452-458.

Mascalchi, M., Lolli, F., Della, N.R. *et al.*, 2004. Huntington disease: volumetric, diffusion-weighted, and magnetization transfer MR imaging of brain. Radiology 232: 867-873.

McKenna P, Warrington EK, 1983. The Graded Naming Test. Windsor, UK: NFER-Nelson.

Mechelli, A., Price, C.J., Friston, K.J., Ashburner, J., 2005. Voxel-based morphometry of the human brain: methods and applications. Current Medical Imaging Reviews 1: 1-9.

Milders, M., Crawford, J.R., Lamb, A., Simpson, S.A., 2003. Differential deficits in expression recognition in gene-carriers and patients with Huntington's disease. Neuropsychologia 41: 1484-1492.

Mitchell, I.J., Heims, H., Neville, E.A., Rickards, H., 2005. Huntington's disease patients show impaired perception of disgust in the gustatory and olfactory modalities. J Neuropsychiatry Clin Neurosci 17: 119-121.

Montagne, B., Kessels, R.P., Kammers, M.P. *et al.*, 2006. Perception of Emotional Facial Expressions at Different Intensities in Early-Symptomatic Huntington's Disease. Eur Neurol 55: 151-154.

Montoya, A., Pelletier, M., Menear, M., Duplessis, E., Richer, F., Lepage, M., 2006. Episodic memory impairment in Huntington's disease: a meta-analysis. Neuropsychologia 44: 1984-1994.

Mühlau, M., Gaser, C., Wohlschlager, A.M. *et al.*, 2007a. Striatal gray matter loss in Huntington's disease is leftward biased. Mov Disord 22: 1169-1173.

Mühlau, M., Weindl, A., Wohlschlager, A.M. *et al.*, 2007b. Voxel-based morphometry indicates relative preservation of the limbic prefrontal cortex in early Huntington disease. J Neural Transm 114: 367-372.

Murphy, F.C., Nimmo-Smith, I., Lawrence, A.D., 2003. Functional neuroanatomy of emotions: a meta-analysis. Cogn Affect Behav Neurosci 3: 207-233.

Nehl, C., Paulsen, J.S., 2004. Cognitive and psychiatric aspects of Huntington disease contribute to functional capacity. J Nerv Ment Dis 192: 72-74.

Nelson HE, Willison J, 1991. The National Adult Reading Test. Windsor, UK: NFER-Nelson.

Nopoulos, P., Magnotta, V.A., Mikos, A., Paulson, H., Andreasen, N.C., Paulsen, J.S., 2007. Morphology of the cerebral cortex in preclinical Huntington's disease. Am J Psychiatry 164: 1428-1434.

Parent, A., Hazrati, L.N., 1995. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. Brain Res Brain Res Rev 20: 91-127.

Paulsen, J.S., Hayden, M., Stout, J.C. *et al.*, 2006a. Preparing for preventive clinical trials: the Predict-HD study. Arch Neurol 63: 883-890.

Paulsen, J.S., Langbehn, D.R., Stout, J.C. *et al.*, 2007. Detection of Huntington's disease decades before diagnosis: The Predict HD study. J Neurol Neurosurg Psychiatry.

Paulsen, J.S., Magnotta, V.A., Mikos, A.E. *et al.*, 2006b. Brain structure in preclinical Huntington's disease. Biol Psychiatry 59: 57-63.

Paulsen, J.S., Zhao, H., Stout, J.C. *et al.*, 2001. Clinical markers of early disease in persons near onset of Huntington's disease. Neurol 57: 658-662.

Paulsen, J.S., Zimbelman, J.L., Hinton, S.C. *et al.*, 2004. fMRI biomarker of early neuronal dysfunction in presymptomatic Huntington's Disease. AJNR Am J Neuroradiol 25: 1715-1721.

Pavese, N., Andrews, T.C., Brooks, D.J. *et al.*, 2003. Progressive striatal and cortical dopamine receptor dysfunction in Huntington's disease: a PET study. Brain 126: 1127-1135.

Peinemann, A., Schuller, S., Pohl, C., Jahn, T., Weindl, A., Kassubek, J., 2005. Executive dysfunction in early stages of Huntington's disease is associated with striatal and insular atrophy: a neuropsychological and voxel-based morphometric study. J Neurol Sci 239: 11-19.

Penney, J.B., Jr., Vonsattel, J.P., MacDonald, M.E., Gusella, J.F., Myers, R.H., 1997. CAG repeat number governs the development rate of pathology in Huntington's disease. Ann Neurol 41: 689-692.

Penney, J.B., Jr., Young, A.B., 1983. Speculations on the functional anatomy of basal ganglia disorders. Annu Rev Neurosci 6: 73-94.

Peterson, B.S., Riddle, M.A., Cohen, D.J., Katz, L.D., Smith, J.C., Leckman, J.F., 1993. Human basal ganglia volume asymmetries on magnetic resonance images. Magn Reson Imaging 11: 493-498.

Phillips, M.L., Young, A.W., Scott, S.K. *et al.*, 1998. Neural responses to facial and vocal expressions of fear and disgust. Proc Biol Sci 265: 1809-1817.

Phillips, M.L., Young, A.W., Senior, C. *et al.*, 1997. A specific neural substrate for perceiving facial expressions of disgust. Nature 389: 495-498.

Poldrack, R.A., 2002. Neural systems for perceptual skill learning. Behav Cogn Neurosci Rev 1: 76-83.

Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., Nichols, T.E., 2007. Guidelines for reporting an fMRI study. Neuroimage 40: 409-414.

Rabiner, L.R., Schafer, R.W., Rader, C.M., 1969. The Chirp z-Transform algorithm and its application. Bell System Technical Journal 48: 1249-1292.

Raz, N., Torres, I.J., Acker, J.D., 1995. Age, gender, and hemispheric differences in human striatum: a quantitative review and new data from in vivo MRI morphometry. Neurobiol Learn Mem 63: 133-142.

Reading, S.A., Dziorny, A.C., Peroutka, L.A. *et al.*, 2004. Functional brain changes in presymptomatic Huntington's disease. Ann Neurol 55: 879-883.

Reading, S.A., Yassa, M.A., Bakker, A. *et al.*, 2005. Regional white matter change in pre-symptomatic Huntington's disease: a diffusion tensor imaging study. Psychiatry Res 140: 55-62.

Reitan RM, Wolfson D, 2004. Comprehensive Handbook of Psychological Assessment. Volume 1. Intellectual and Neuropsychological Assessment. Hoboken, NJ: John Wiley and Sons.

Rich, J.B., Troyer, A.K., Bylsma, F.W., Brandt, J., 1999. Longitudinal analysis of phonemic clustering and switching during word-list generation in Huntington's disease. Neuropsychology 13: 525-531.

Ridgway, G.R., Henley, S.M., Rohrer, J.D., Scahill, R.I., Warren, J.D., Fox, N.C., 2008. Ten simple rules for reporting voxel-based morphometry studies. Neuroimage 40: 1429-1435.

Rosas, H.D., Goodman, J., Chen, Y.I. *et al.*, 2001. Striatal volume loss in HD as measured by MRI and the influence of CAG repeat. Neurol 57: 1025-1028.

Rosas, H.D., Hevelone, N.D., Zaleta, A.K., Greve, D.N., Salat, D.H., Fischl, B., 2005. Regional cortical thinning in preclinical Huntington disease and its relationship to cognition. Neurol 65: 745-747.

Rosas, H.D., Koroshetz, W.J., Chen, Y.I. *et al.*, 2003. Evidence for more widespread cerebral pathology in early HD: an MRI-based morphometric analysis. Neurol 60: 1615-1620.

Rosas, H.D., Liu, A.K., Hersch, S. *et al.*, 2002. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurol 58: 695-701.

Rosas, H.D., Salat, D.H., Lee, S.Y. *et al.*, 2008. Cerebral cortex and the clinical expression of Huntington's disease: complexity and heterogeneity. Brain 131: 1057-1068.

Rosas, H.D., Tuch, D.S., Hevelone, N.D. *et al.*, 2006. Diffusion tensor imaging in presymptomatic and early Huntington's disease: Selective white matter pathology and its relationship to clinical measures. Mov Disord 21: 1317-1325.

Rosenblatt, A., Liang, K.Y., Zhou, H. *et al.*, 2006. The association of CAG repeat length with clinical progression in Huntington disease. Neurol 66: 1016-1020.

Rosenblatt, A., Margolis, R.L., Becher, M.W. *et al.*, 1998. Does CAG repeat number predict the rate of pathological changes in Huntington's disease? Ann Neurol 44: 708-709.

Rothman, K.J., 1990. No adjustments are needed for multiple comparisons. Epidemiology 1: 43-46.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63: 581-92.

Rubinsztein, D.C., Leggo, J., Coles, R. *et al.*, 1996. Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. Am J Hum Genet 59: 16-22.

Ruocco, H.H., Bonilha, L., Li, L.M., Lopes-Cendes, I., Cendes, F., 2007. Longitudinal analysis of regional gray matter loss in Huntington disease: effects of the length of the CAG repeat. J Neurol Neurosurg Psychiatry 79: 130-135.

Ruocco, H.H., Lopes-Cendes, I., Li, L.M., Santos-Silva, M., Cendes, F., 2006. Striatal and extrastriatal atrophy in Huntington's disease and its relationship with length of the CAG repeat. Braz J Med Biol Res 39: 1129-1136.

Sahin, B., Acer, N., Sonmez, O.F. *et al.*, 2007. Comparison of four methods for the estimation of intracranial volume: a gold standard study. Clin Anat 20: 766-773.

Saling, L.L., Phillips, J.G., 2007. Automatic behaviour: efficient not mindless. Brain Res Bull 73: 1-20.

Samejima, K., Doya, K., 2007. Multiple representations of belief states and action values in corticobasal ganglia loops. Ann N Y Acad Sci 1104: 213-228.

Scahill, R.I., Frost, C., Jenkins, R., Whitwell, J.L., Rossor, M.N., Fox, N.C., 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch Neurol 60: 989-994.

Scott, S.K., Young, A.W., Calder, A.J., Hellawell, D.J., Aggleton, J.P., Johnson, M., 1997. Impaired auditory recognition of fear and anger following bilateral amygdala lesions. Nature 385: 254-257.

Shelbourne, P.F., Keller-McGandy, C., Bi, W.L. *et al.*, 2007. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Hum Mol Genet 16: 1133-1142.

Shoulson, I., 1981. Huntington disease: functional capacities in patients treated with neuroleptic and antidepressant drugs. Neurol 31: 1333-1335.

Shoulson, I., Fahn, S., 1979. Huntington disease: clinical care and evaluation. Neurol 29: 1-3.

Sieradzan, K., Mann, D.M., Dodge, A., 1997. Clinical presentation and patterns of regional cerebral atrophy related to the length of trinucleotide repeat expansion in patients with adult onset Huntington's disease. Neurosci Lett 225: 45-48.

Simmons, A., Matthews, S.C., Stein, M.B., Paulus, M.P., 2004. Anticipation of emotionally aversive visual stimuli activates right insula. Neuroreport 15: 2261-2265.

Simon, D., Craig, K.D., Gosselin, F., Belin, P., Rainville, P., 2007. Recognition and discrimination of prototypical dynamic expressions of pain and emotions. Pain 135: 55-64.

Skuse, D., Morris, J., Lawrence, K., 2003. The amygdala and development of the social brain. Ann N Y Acad Sci 1008: 91-101.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17: 87-97.

Smith, A., 1968. The Symbol Digit Modalities Test: a neuropsychologic test for economic screening of learning and other cerebral disorders. Learning Disorders 3: 83-91.

Snowden, J., Craufurd, D., Griffiths, H., Thompson, J., Neary, D., 2001. Longitudinal evaluation of cognitive disorder in Huntington's disease. J Int Neuropsychol Soc 7: 33-44.

Snowden, J.S., Craufurd, D., Thompson, J., Neary, D., 2002. Psychomotor, executive, and memory function in preclinical Huntington's disease. J Clin Exp Neuropsychol 24: 133-145.

Solomon, A.C., Stout, J.C., Johnson, S.A. *et al.*, 2007. Verbal episodic memory declines prior to diagnosis in Huntington's disease. Neuropsychologia 45: 1767-1776.

Sprengelmeyer, R., Rausch, M., Eysel, U.T., Przuntek, H., 1998. Neural structures associated with recognition of facial expressions of basic emotions. Proc Biol Sci 265: 1927-1931.

Sprengelmeyer, R., Schroeder, U., Young, A.W., Epplen, J.T., 2005. Disgust in pre-clinical Huntington's disease: A longitudinal study. Neuropsychologia 44: 518-533.

Sprengelmeyer, R., Young, A.W., Calder, A.J. *et al.*, 1996. Loss of disgust - perception of faces and emotions in Huntington's disease. Brain 119: 1647-1665.

Sprengelmeyer, R., Young, A.W., Sprengelmeyer, A. *et al.*, 1997. Recognition of facial expressions: Selective impairment of specific emotions in Huntington's disease. Cognitive Neuropsychology 14: 839-879.

Squitieri, F., Cannella, M., Simonelli, M., 2002. CAG mutation effect on rate of progression in Huntington's disease. Neurol Sci 23 Suppl 2: S107-S108.

Starkstein, S.E., Brandt, J., Bylsma, F., Peyser, C., Folstein, M., Folstein, S.E., 1992. Neuropsychological correlates of brain atrophy in Huntington's disease: a magnetic resonance imaging study. Neuroradiology 34: 487-489.

Stout, J.C., Weaver, M., Solomon, A.C. *et al.*, 2007. Are cognitive changes progressive in prediagnostic HD? Cogn Behav Neurol 20: 212-218.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18: 643-662.

Thieben, M.J., Duggins, A.J., Good, C.D. *et al.*, 2002. The distribution of structural neuropathology in pre-clinical Huntington's disease. Brain 125: 1815-1828.

Thiruvady, D., Georgiou-Karistianis, N., Egan, G. *et al.*, 2006. Functional connectivity of the prefrontal cortex in Huntington's disease. J Neurol Neurosurg Psychiatry 78: 127-133.

Thompson, J.C., Snowden, J.S., Craufurd, D., Neary, D., 2002. Behavior in Huntington's disease: dissociating cognition-based and mood-based changes. J Neuropsychiatry Clin Neurosci 14: 37-43.

Tibben A. Genetic counselling and presymptomatic testing. In: Bates G, Harper P, Jones L, editors. Huntington's Disease. Oxford: Oxford University Press; 2002. p. 198-248.

Troyer, A.K., Moscovitch, M., Winocur, G., Alexander, M.P., Stuss, D., 1998. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. Neuropsychologia 36: 499-504.

van Oostrom, J.C., Maguire, R.P., Verschuuren-Bemelmans, C.C. *et al.*, 2005. Striatal dopamine D2 receptors, metabolism, and volume in preclinical Huntington disease. Neurol 65: 941-943.

Verny, C., Allain, P., Prudean, A. *et al.*, 2007. Cognitive changes in asymptomatic carriers of the Huntington disease mutation gene. Eur J Neurol 14: 1344-1350.

Vonsattel, J.P., Myers, R.H., Stevens, T.J., Ferrante, R.J., Bird, E.D., Richardson, E.P., Jr., 1985. Neuropathological classification of Huntington's disease. J Neuropathol Exp Neurol 44: 559-577.

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans Med Imaging 26: 405-421.

Wang, K., Hoosain, R., Yang, R.M., Meng, Y., Wang, C.Q., 2003. Impairment of recognition of disgust in Chinese with Huntington's or Wilson's disease. Neuropsychologia 41: 527-537.

Wanker E, Droge A. Structural biology of Huntington's disease. In: Bates G, Harper P, Jones L, editors. Huntington's Disease. Oxford: Oxford University Press; 2002. p. 327-47.

Ward, J., Sheppard, J.M., Shpritz, B., Margolis, R.L., Rosenblatt, A., Brandt, J., 2006. A four-year prospective study of cognitive functioning in Huntington's disease. J Int Neuropsychol Soc 12: 445-454.

Warrington EK, 1984. Recognition Memory Test. Windsor, UK: NFER-Nelson.

Warrington, E.K., 2000. Homophone meaning generation: a new test of verbal switching for the detection of frontal lobe dysfunction. J Int Neuropsychol Soc 6: 643-648.

Warrington EK, James M, 1991. The visual object and space perception battery. Bury St Edmunds: Thames Valley Test Co.

Wechsler DA, 1981. Wechsler Adult Intelligence Scale-Revised. New York: The Psychological Corporation.

Whitwell, J.L., Crum, W.R., Watt, H.C., Fox, N.C., 2001. Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging. AJNR Am J Neuroradiol 22: 1483-1489.

Whitwell, J.L., Schott, J.M., Lewis, E.B., MacManus, D.G., Fox, N.C., 2004. Using nine degrees-of-freedom registration to correct for changes in voxel size in serial MRI studies. Magn Reson Imaging 22: 993-999.

Williams, J.K., Hamilton, R., Nehl, C. *et al.*, 2007. "No one else sees the difference: "family members' perceptions of changes in persons with preclinical Huntington disease. Am J Med Genet B Neuropsychiatr Genet 144: 636-641.

Witjes-Ane, M.N., Mertens, B., van Vugt, J.P., Bachoud-Levi, A.C., van Ommen, G.J., Roos, R.A., 2007. Longitudinal evaluation of "presymptomatic" carriers of Huntington's disease. J Neuropsychiatry Clin Neurosci 19: 310-317.

Witjes-Ane, M.N., Vegter-van der Vlis, M., van Vugt, J.P. *et al.*, 2003. Cognitive and motor functioning in gene carriers for Huntington's disease: a baseline study. J Neuropsychiatry Clin Neurosci 15: 7-16.

Wolf, R.C., Vasic, N., Schonfeldt-Lecuona, C., Ecker, D., Landwehrmeyer, G.B., 2008. Cortical dysfunction in patients with Huntington's disease during working memory performance. Hum Brain Mapp.

Wolf, R.C., Vasic, N., Schonfeldt-Lecuona, C., Landwehrmeyer, G.B., Ecker, D., 2007. Dorsolateral prefrontal cortex dysfunction in presymptomatic Huntington's disease: evidence from event-related fMRI. Brain 130: 2845-2857.

Wong, B., Cronin-Golomb, A., Neargarder, S., 2005. Patterns of visual scanning as predictors of emotion identification in normal aging. Neuropsychology 19: 739-749.

Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C., 1998. Automated image registration: I. General methods and intrasubject, intramodality validation. J Comput Assist Tomogr 22: 139-152.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J Cereb Blood Flow Metab 12: 900-918.

Zakzanis, K.K., Mraz, R., Graham, S.J., 2005. An fMRI study of the Trail Making Test. Neuropsychologia 43: 1878-1886.

Zimbelman, J.L., Paulsen, J.S., Mikos, A., Reynolds, N.C., Hoffmann, R.G., Rao, S.M., 2007. fMRI detection of early neural dysfunction in preclinical Huntington's disease. J Int Neuropsychol Soc 13: 758-769.