Structural cortical grey matter changes during the transition from premanifest to manifest Huntington's disease: A methodological evaluation

Thesis submitted for the degree of Doctor of Philosophy

Eileanoir Johnson

Institute of Neurology

University College London

2018

Declaration

I, Eileanoir Johnson, 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

Huntington's disease (HD) is a genetic neurodegenerative disease characterised by motor, cognitive and psychiatric symptoms. Atrophy of subcortical brain structures has been well characterised and changes in the white matter are being mapped with increasing frequency, but structural changes in the cortex have been relatively overlooked in previous research. With recently trialled therapies specifically targeting the cortex, a better understanding of the pattern and progression of atrophy in this region should provide valuable measures for determining the impact of these novel treatments on the degenerative process.

This thesis performs a methodological comparison aimed at optimising techniques to measure cortical characteristics in an HD cohort, and then applies the optimised techniques in a group of HD gene carriers undergoing conversion from pre-manifest HD to manifest HD quantifying cortical change during this critical period. Several tools for the quantification of cortical volume and cortical thickness are examined via detailed analyses using two datasets. This investigation results in a series of recommendations for the use of such tools, as well as the selection of the most appropriate measures for use in the second part of this thesis. In addition, since subcortical atrophy measures are widely used in HD research, the performance of one of the segmentation tools was evaluated by comparison with manual segmentations of the caudate and putamen. Finally, a novel multivariate analysis method is applied, based on the principles of DCM, to measure the rate, timing and acceleration of cortical GM change in a subgroup of 49 motor converters from the TRACK-HD cohort. This cortical atrophy is then related both to the biological underpinning of the disease in terms of CAG length and also the behavioural presentation of motor and cognitive symptoms. These findings present the first detailed characterisation of cortical grey matter change in HD, and have important implications for the understanding of HD progression.

ACKNOWLEDGEMENTS

Firstly I would like to thank all of the participants who gave their time to contribute to this research. I am so glad that I got to meet many of you during data collection for TrackOn-HD. You are a continuous source of motivation to do the best work that I can do. I would also like to thank the CHDI foundation for funding the TRACK-HD and TrackOn studies and my role in the project.

To everyone at the HD centre: you have all helped me to get this thesis finished, and I am so very lucky to be part of such an incredible team. I would especially like to thank Sarah Tabrizi for leading us and for giving me the opportunity to undertake this PhD. To Elin and Helen, my predecessors, who both provided me with an aim and advice on how to produce this body of work without losing faith and sanity. I would like to thank Nicky Hobbs for not only giving me an insight into the world outside of academia but also for being such an inspiration. Geraint Rees for his astute contribution to a number of chapters, and Gabriel Ziegler and Will Penny for both putting in huge amounts of work to develop the models whilst also trying to understand HD. This thesis would not have been nearly as fascinating without your input.

Special thanks to Sarah Gregory, who has provided the unofficial guidance of a third supervisor along with plenty of biscuits and laughs to see me through. You put in a massive amount of work to guide me through this process and I am so glad to have been able to learn from you. And my heartfelt thanks to my supervisor, Rachael Scahill. Your encouragement and total faith in my ability to complete this PhD has re-energised me when I lost motivation. I am hugely grateful to have a PhD supervisor who I admire so much, professionally and personally, and who has helped to make completing this thesis so enjoyable.

Finally, to my family; my mum and dad who always taught me that I should do what I enjoy and whose encouragement never wavered. My sister, Caragh, whose constant positivity and reassuring words have boosted me throughout. And to Mischa. Your late-night words helped to calm me when I felt overwhelmed and your dogged belief in me pushed me on through the lows. Thank you.

CONTRIBUTORS

I would like to thank everyone who contributed to this thesis, but particularly those people listed here for their individual contributions.

All chapters:

Rachael Scahill and Sarah Gregory

- Input on study design and proof reading

Track-HD, TrackOn-HD and PADDINGTON study teams

- Data collection and processing for TRACK-HD, TrackOn-HD and PADDINGTON. My personal contribution to data collection and processing is listed in sections 2.1.3.1 and 2.3.3.6.

Chapter 5:

Elin Rees

- Segmentation of the manual putamen ROIs used in this chapter.

Chapter 6:

Gabriel Ziegler and Will Penny

- Development of the sDCM models for this chapter (adapted from Ziegler et al. 2017) and associated Matlab code used to perform the analysis and output the figures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3T	3 Tesla
Advanced Normalization Tools	ANTs
BBSI	Brain boundary shift integral
BET	Brain Extraction Tool
BMS	Bayesian model selection
BRAINS	Brain Research: Analysis of Images, Networks, and Systems
BSI	Boundary shift integral
CAG	Cytosine-Adenosine-Guanine
CAP	CAG age product scaled
CBSI	Caudate boundary shift integral
CGM	Cortical grey matter
CSF	Cerebrospinal fluid
CT	Cortical thickness
DARTEL	Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra
DBS	Disease Burden Score
DCM	Dynamic causal modelling
DTI	Diffusion tensor imaging
FAST	FMRIB's Automated Segmentation Tool
FLIRT	FMRIB's Linear Image Registration Tool
fMRI	Functional MRI
FMRIB	Functional Magnetic Resonance Imaging of the Brain - Analysis Group, Oxford University, UK
FNIRT	FMRIB's Non-linear Image Registration Tool
FOV	Field of view
FSL	FMRIB Software Library
FWE	Family-Wise Error
GLM	General linear modelling
GM	Grey matter
GT	Ground truth
HD	Huntington's disease
HD1	Huntington’s disease – stage 1
HD2	Huntington’s disease – stage 2
HTT	Huntingtin protein

ICC	Intraclass correlation coefficient
MALP-EM	Multi-Atlas Label Propagation with Expectation–Maximisation based refinement
mHTT	Mutant huntingtin
MIDAS	Medical Image Display and Analysis Software
MNI	Montreal Neurological Institute
MPRAGE	Magnetization prepared rapid gradient echo
MRI	Magnetic resonance imaging
NfL	Neurofilament light protein
PADDINGTON	Pharmacodynamic Approaches to Demonstration of Disease-modification in Huntington's disease by SEN0014196
PGD	Preimplantation genetic diagnosis
PolyQ	Polyglutamine
Pre-HD	Pre-HD
PreHD-A	Pre-manifest Huntington’s disease – further from predicted disease onset
PreHD-B	Pre-manifest Huntington’s disease – closer to predicted disease onset
PVE	Partial volume estimates
QC	Quality control
RF	radiofrequency
ROI	Region of interest
rs-fMRI	Resting state fMRI
SD	Standard deviation
SDMT	Symbol digit modality test
sMRI	Structural MRI
SPM	Statistical Parametric Mapping
TE	Echo time
TFC	Total Functional Capacity
TIV	Total intracranial volume
TMS	Total Motor Score
TR	Repetition time
UCL	University College London
UHDRS	Unified Huntington's Disease Rating Scale
VBM	Voxel-Based Morphometry
WM	White matter
WP2	Work package 2

AIMS OF THIS THESIS

In the period immediately prior to Huntington's Disease (HD) motor diagnosis there is an increase in symptom severity that is thought to be accompanied by an increase in atrophy within the cortical grey matter (CGM) in the brain (Tabrizi et al. 2013). The pattern of CGM atrophy that occurs during the transition from pre-manifest HD (pre-HD) to manifest HD has not been fully characterised and its relationship to the onset of symptoms is currently unclear. Magnetic resonance imaging (MRI) can be used to capture images of the brain in vivo across many stages of HD, and is subsequently a vital tool in the characterization of atrophy in HD. However, it is challenging to accurately quantify cortical change from MRI despite the availability of many software tools designed to achieve this goal. Automated tools are regularly used to analyse a number of cortical features, such as volume and thickness of the grey matter, yet due to a large range of individual differences along with the complicated gyrification within the human cortex these tools often suffer from errors in segmentation that can result in inconsistent findings between studies. In addition, while previous studies have been designed to quantify cortical change in HD the majority of these studies have used mass-univariate linear methods, which are difficult to apply in cohorts with data from variable time points and do not provide information about potential accelerations or decelerations in atrophy across disease stages. These issues combined with a number of other methodological problems have contributed to a relatively limited understanding about the trajectory of cortical change in HD**.** To understand the relationship between neurological change and HD progression, a detailed characterisation of the changes that occur in the cortex during symptom onset is required.

OVERALL AIM

This thesis aims to evaluate and optimise the most accurate tools for quantifying cortical atrophy from MRI scans in a large cohort of gene carriers across the HD disease spectrum and to undertake a characterisation of CGM change during HD motor onset.

SPECIFIC AIMS

- To investigate the most appropriate methods to detect and measure CGM changes in HD and to optimise these techniques resulting in a ready-to-use pipeline for analysis.
- To conduct a detailed characterisation of what CGM changes occur during the period immediately surrounding HD diagnosis.

- To investigate how CGM brain changes relate to the biological underpinning and behavioural manifestation of HD.

1. INTRODUCTION

This chapter will provide an overview of key topics covered by this thesis. An introduction to Huntington's disease (HD) will be provided, followed by a summary of the process of generating MRI scans. Different MRI analysis techniques will be discussed, and an overview of findings from previous work studying neural atrophy in HD using structural MRI (sMRI) will be reviewed. Finally, an outline of the current thesis will be provided within the context of previous literature.

1.1. Huntington's disease

HD is a genetic neurodegenerative disease characterised by motor, cognitive and psychiatric symptoms. The disease is caused by an expanded CAG repeat in the *HTT* gene and it is fully penetrant and incurable (Tabrizi et al. 2012; Tabrizi et al. 2013). An individual can be identified as gene positive many years before the onset of symptoms but the formal diagnosis is only made in the presence of motor symptoms via use of the Unified Huntington's Disease Rating Scale (UHDRS; Huntington Study Group 1996).

HD is caused by an expanded CAG trinucleotide repeated in *HTT*, the gene responsible for encoding the protein huntingtin. The expanded CAG repeat results in the production of a mutant form of huntingtin that has an unusually long polyglutamine (polyQ) sequence (Bates et al. 2015). Mutant huntingtin (mHTT) exhibits toxic properties that cause cellular dysfunction and neuronal death. The medium spiny neurons found in the striatum are the most vulnerable to damage by mHTT, however neurons in other brain regions are also damaged by mHTT, leading to widespread neural atrophy over the course of HD. While HD has typically been viewed as a disease affecting the central nervous system, it is now understood to affect the whole body, with symptoms such as weight loss, muscle changes and changes to the peripheral immune system commonly occurring in patients alongside neurological symptoms (Björkqvist et al. 2008; Carroll et al. 2015).

1.1.1. Mechanisms and pathophysiology

Within the *HTT* gene is a CAG repeat that has varying length in the normal population. The length of this CAG repeat can be measured in any individual, with the length determining penetrance of HD. The healthy length of CAG repeats is between 6-35 repeats. For individuals with ≥40 repeats the disease will be fully penetrant. Repeat lengths of 36-39 are classed as

reduced penetrance, with some individuals in this range showing no signs of HD throughout their lives and others being clinically diagnosed as having HD, although generally later in life.

HD is autosomal dominant, meaning that if one parent is affected each child has a 50% chance of inheriting the expanded CAG repeat. In addition, the CAG repeat shows instability between parents and children, with longer CAG repeats often also having greater instability. Between generations the repeat can typically increase or decrease by a small number of CAGs, although large increases in CAG repeat length are sometimes seen; in some cases this may result in the appearance of a new mutation with an unaffected parent passing on a CAG repeat length in the pathogenic range. CAG repeat length not only predicts whether an individual will develop HD, but also predicts the age at which symptom onset occurs. Within the most common CAG range measured in individuals with HD, 40-55 repeats, approximately 56% of the variation in age at motor onset is accounted for by CAG length. Of the remaining variation, a large percentage (38-56%) can be attributed to genetic modifiers of onset and progression, although little is currently known about genetic modifiers in HD (Moss et al. 2017).

The normal form of huntingtin is expressed throughout the whole body, with the level of expression varying across different regions of the body. The function of huntingtin is not well understood, however it is believed to be important for the development of the nervous system, for cell adhesion and is understood to affect brain-derived neurotropic factor (BDNF), a protein supporting neuronal survival and growth (Bates et al. 2015). Currently, a detailed understanding of how huntingtin affects these processes is lacking, but it is hypothesised that the polyglutamine expansion results in the loss or modulation of normal huntingtin functioning and that a critical level of this abnormal huntingtin causes enough damage to cause symptom onset.

1.1.2. Epidemiology

Within Western populations, the prevalence of HD is thought to be between 10.6-13.7 individuals per 100,000 (Fisher & Hayden 2014; Evans et al. 2013; Morrison et al. 2011), with higher prevalence in individuals with European ancestry. In Asian and African cultures, there are typically lower rates of HD (diagnosis seems to occur around one-tenth as frequently), however often the results of epidemiology studies in these regions are based on case studies and thus the exact prevalence is unclear. There have also been pockets of extremely high prevalence reported, with the most widely studied being the Venezuelan cohort in Maracaibo which led to the discovery of the HD gene (The U.S.-Venezuela Collaborative et al. 2004).

Over the last few decades, the rates of HD appear to have increased (Bates et al. 2015). This is likely to be due to the development and subsequent increasing availability of a genetic test for HD. Estimations of prevalence described before the development of genetic testing possibly underestimated sporadic cases of HD (5-8% of cases). Furthermore, the genetic test can now confirm late-onset HD, which may have previously been underestimated due to concurrent aging or other related factors. Finally, longer lifespans are also probable contributors to increasing rates of prevalence, with longer lifespans resulting in more late-onset cases of HD.

1.1.3. Diagnosis and symptomology

The official clinical diagnosis threshold for HD is based on motor symptoms, however there is a long pre-HD phase whereby subtle symptoms of the disease begin to emerge. This phase can last for 10-15 years before motor diagnosis and includes psychiatric disturbances such as depression, anxiety or aggression, and cognitive symptoms including worsening short-term memory and attentional deficits. During this phase, subtle and progressively worsening motor symptoms are typically experienced. Onset of the pre-HD phase is related to CAG length, but typically starts during an individual's late thirties to mid-forties.

The initial symptoms of HD experienced during this phase are highly variable between patients (Papp et al. 2011), but as the disease progresses symptoms become more consistent; with visuomotor performance and working memory largely affected in late premanifest HD and motor symptoms beginning to become more obvious (Papp et al. 2011). Diagnosis of HD occurs in the presence of irrefutable motor symptoms (as defined by the UHDRS) and a confirmed family history of HD or positive genetic test (Huntington Study Group 1996). The UHDRS rating scale ranges from 0 to 4, with 0 indicating no motor abnormalities suggestive of HD, and 4 indicating motor symptoms that are ≥99% likely to be due to HD. A score of 4 defines motor onset, and symbolises the period known as 'manifest' HD.

After onset, motor, cognitive and behavioural symptoms continue to advance in a variable pattern until death. In addition to being associated with earlier age at onset, longer CAG repeat length is associated with faster progression once symptoms have begun, although this relationship is not as robust as that between CAG length and onset (Ross et al. 2014).

The discovery of the HD gene was made in 1993 (MacDonald et al. 1993), and since then genetic testing has been used as both a diagnostic tool and a predictive test in HD. If a patient comes to a clinic with symptoms that are characteristic of HD, a genetic test might be used to diagnose or exclude HD as a cause of these symptoms. Alternatively, genetic testing

can be performed as a predictive test many years before onset. The implications of receiving a positive genetic test result are significant, and so in 1994 a set of international guidelines were developed to reduce negative psychological and physical harm that may be associated with a positive result (Anon 1994). These guidelines recommended that in order to undergo predictive testing an individual should be given genetic counselling and a psychological assessment, along with a neurological examination and time to reconsider whether they definitely want the test performed. In addition, the guidelines stipulate that results should be provided in person with the opportunity to receive post-test support. Children under 18 should not be tested unless they are symptomatic. These recommendations were recently updated and now include advice on the provision of information about preimplantation genetic diagnosis (PGD; MacLeod et al. 2013). The rate of testing in individuals at risk from HD varies, ranging from 3-4% in Germany, Austria and Switzerland to 24% in Denmark. The rate of predictive testing in the UK is reported to be around 18% (Tibben 2007).

1.1.3.1. Motor symptoms

Motor symptoms are the most widely recognised and visible signs of HD. They generally fall into two broad categories. The first category is involuntary movement, with chorea being the most common motor symptom of HD (Bates et al. 2015). Chorea is characterised by quick, irregular and unpredictable movements, often of the limbs. The second category of motor symptoms is the impairment of voluntary movements, which includes lack of coordination and bradykinesia. Bradykinesia refers to the slowing of movement, and encompasses an inability to move the body quickly. Bradykinesia is typically associated with long CAG repeats, and is especially prevalent in juvenile HD, although it also occurs in the later stages of typical HD progression. Motor symptoms are slowly progressing in most individuals, and are measured using the Total Motor Score section of the UHDRS (Huntington Study Group 1996). This scale examines different features of motor disturbance in HD, including chorea, bradykinesia, speech, eye movements, gait and other features.

1.1.3.2. Cognitive symptoms

Cognitive symptoms are another key element of HD progression. Similarly to motor progression, changes in cognition occur slowly. Subtle cognitive symptoms can be seen around 10 years prior to clinical diagnosis, with the earliest cognitive signs including problems with visuomotor integration, psychomotor speed, emotion recognition and executive functioning (Bates et al. 2015). As a patient progresses, problems are also experienced with attention,

mental flexibility, learning, episodic and working memory along with more general slowing in cognition. Semantic memory, spatial awareness and language do not show significant deterioration as in other neurodegenerative conditions, such as Alzheimer's.

1.1.3.3. Psychiatric symptoms

The psychiatric symptoms associated with HD are much more variable than both motor and cognitive symptoms (Bates et al. 2015). Depression is commonly experienced in HD gene carriers, both before and after disease onset. In addition, apathy and irritability are both seen frequently in HD. Apathy increases with increasing disease progression, getting increasingly debilitating.

1.1.3.4. Impact of Huntington's disease on quality of life

As HD progresses, quality of life patients is reduced for many by the increasing severity of HD symptoms (Read et al. 2013). Depressive mood and decreases in functional ability (ability to work, feed oneself, cook etc.) have been shown as two of the greatest contributors to reduced quality of life (Ho et al. 2009), with apathy also being been linked to worse quality of life in HD (Ready et al. 2008). By end-stage disease, patients are incapacitated and are often receiving care in a specialised care facility. Treatments should aim to improve quality of life where possible.

1.1.4. Models of HD onset and progression

Due to the strong relationship between age and CAG, these two variables can be used to predict estimated disease onset and progression, and to categorise participants on their disease severity and exposure to the mutant form of huntingtin. A number of models exist to scale participants on these factors, with three commonly used models described here.

1.1.4.1. Disease-burden score

One of the commonly used models is the disease-burden score. This model was developed by Penney et al. (1997) and is based on the relationship between post-mortem striatal atrophy, age at death and CAG repeat length. Their study found that these three factors were linearly related and had an intercept at 35.5, and concluded that 35.5 was the largest CAG repeat whereby no pathology would develop in the striatum. They also hypothesised that the pathological process would develop from birth in a linear way. Thus they modelled: disease burden = (CAG – 35.5) x current age. This approximation of disease burden

is frequently used as a theoretical model in HD studies, including the TRACK-HD and TrackOn-HD studies to scale participants based on their approximate exposure to the mutant huntingtin protein and thus approximate disease stage. However, the model was based on post-mortem data and is thus biased towards end-stage disease, and assumes that atrophy progresses linearly from birth. This model also does not take into account other environmental or genetic factors, which may influence the progression and onset of HD.

1.1.4.2. Predicted years-to-onset

In addition, there are models of estimated disease onset. The most widely used is that developed by Langbehn et al. (2004). This model is a parametric survival model that estimates how far a participant is from clinical onset. Data from approximately 3000 pre-HD and manifest participants were used to develop the model. The calculation uses a conditional probability model, which means that given an individual's CAG repeat and age, and given that motor onset has not yet occurred, the probability that they will undergo motor onset by a given age can be calculated. A strength of the Langbehn model is that by incorporating a large number of both pre-HD and manifest participants selection biases are reduced.

1.1.4.3. CAG age product scaled

The CAG age product scaled (CAP) is a more recent method of estimating disease stage. The CAP score was developed on PREDICT-HD data, and validated using a longitudinal receiver operating characteristic analysis to show that it was a strong predictor of onset, especially if onset was within two years of calculation (Zhang et al. 2011). CAP score is often used as a normalised measure of onset, and can be used across pre-HD and manifest participants. Figure 1.1 shows a schematic representing the hypothesised course of HD symptoms with age, normalised CAP score and motor diagnosis (Ross et al. 2014).

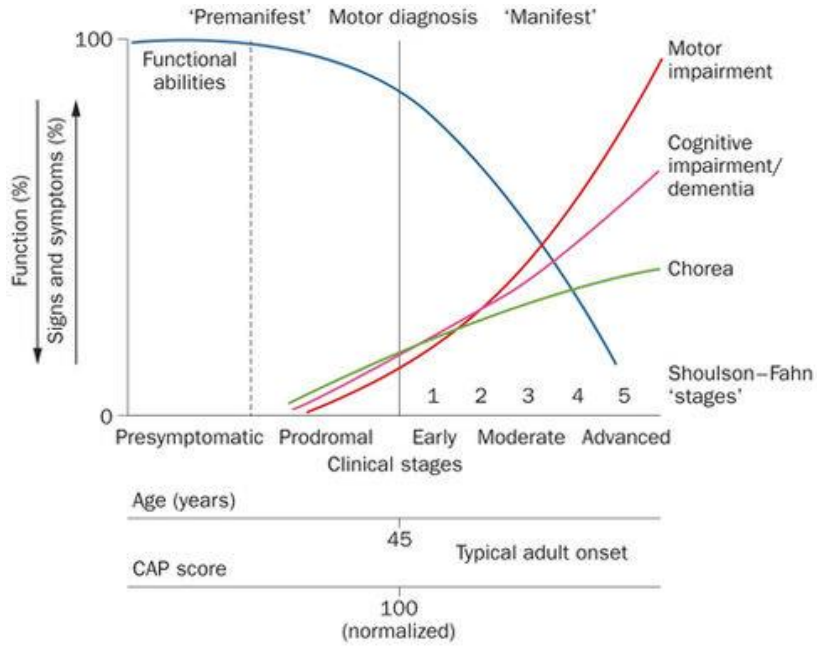


Figure 1.1. A schematic representation of the course of HD, showing the typical adult onset and normalised CAP scores vs. symptom onset. Adapted from Ross et al. (2014).

1.1.5. Management and therapeutic trials

There are currently no disease-modifying treatments for HD. Instead, clinicians focus on managing the symptoms of HD. There are two drugs specifically approved for the treatment of chorea in HD: tetrabenazine and deutetrabenazine, which show a moderate effect for reducing chorea (Rodrigues & Wild 2017). In addition to these drugs, commonly used psychiatric drugs are often prescribed to reduce psychiatric symptoms associated with HD.

Further to treatment with drugs, the management of HD can involve a wide range of health professionals since the disease has an impact on many aspects of a patient's life. In addition to neurologists and psychiatrists or psychologists, a patient may also require care from occupational health practitioners and physiotherapists to provide help in adjusting to HD-related movement or balance issues.

There have been almost 100 clinical trials aimed at testing disease-modifying treatments for HD, none of which have been successful (Rodrigues & Wild 2017). There are also a number of on-going clinical trials in HD. Some of the most promising approaches include targeting the transcription of HTT or the translation of HTT mRNA (Wild & Tabrizi 2017). In 2016 a landmark clinical trial began aimed at reducing the production of mutant huntingtin by targeting the pre-mRNA transcript of the *HTT* gene. The trial was phase IB/IIA with the initial aims being safety and tolerability. Following positive safety results, the study has been rolled into an open label extension trial, beginning in late 2017. This is the first trial aimed at this type

of intervention, however more trials testing different techniques of HTT lowering are due to begin soon.

1.1.6. Post-mortem results in HD

The neural atrophy that occurs over the course of HD has been characterised via a number of pathological studies measuring the brains of HD patients post-mortem (Vonsattel et al. 1985; De La Monte et al. 1988; Mann et al. 1993). While even the earliest post-mortem studies report that the most striking feature of these brains is atrophy of the caudate and putamen, there is also atrophy of the white matter (WM) and cortex noted in most studies. Figure 1.2. shows an example of a post-mortem brain damaged from HD. One study reported that the volume of the caudate was reduced by 53% and the putamen by 46% in end-stage HD patients compared to controls. This study also showed a reduction of the WM (13%), and relatively uniform atrophy of the cortex, with overall cortical volume reduced by 23% compared to controls, and only the temporal lobes being relatively spared from atrophy (Halliday et al. 1998). Within the cortex, the occipital lobe showed the greatest difference compared to controls (28%), frontal and parietal regions were reduced by 22%, and the temporal lobe was 17% lower than in control participants. This study also reported that the extent of atrophy present in the cortex, but not the subcortical regions, correlated with CAG repeat length (Halliday et al. 1998).

More recently, the cell loss in the motor cortex was associated with level of motor dysfunction in HD (Thu et al. 2010). The same study also found that patients rated as having more severe mood symptoms had greater cellular loss in the cingulate cortex, a region known to process emotion, indicating a link between cortical atrophy and HD symptomology.

While the majority of post-mortem results are from patients who are at end-stage disease, they are indicative of widespread volume loss that occurs over a number of years extending beyond the striatum. Due to the small number of post-mortem brains available from other stages of HD, it is only via in-vivo techniques including MRI that we can measure the progression of atrophy and understand changes occurring in earlier phases of the disease.

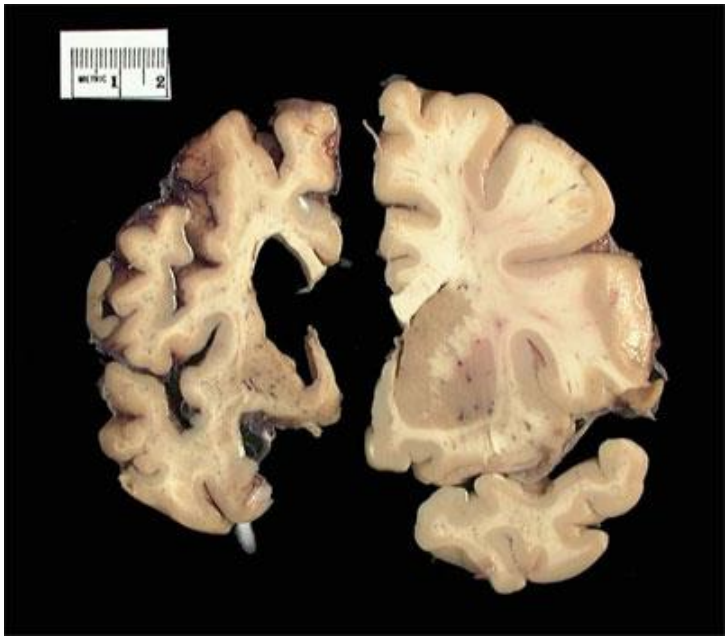


Figure 1.2. An example of a post-mortem HD brain, left, compared to a healthy brain, right. Adapted from https://hbtrc.mclean.harvard.edu/about/tour/slideview.php?page=41.

1.1.1. Biomarkers in HD

In order to quantify the effects of therapeutic interventions, the validation of accurate and sensitive biomarkers is essential. A biomarker is a measureable indicator of a medical state or disease. It should be objective, easy to measure and understand, quantifiable, sensitive and preferably non-invasive (Paulsen 2009). The TRACK-HD study was designed to test a range of potential biomarkers in HD (Tabrizi et al. 2009). By collecting cognitive, psychological, motor and imaging data for controls, pre-HD and manifest HD participants over 4 years the best markers of disease progression at different stages of HD could be selected.

In the TRACK-HD analysis conducted after 24 months of data collection the effect sizes of potential biomarkers were compared in a large cohort which included 123 controls, 120 pre-HD and 123 early HD participants at baseline (Tabrizi et al. 2012). Change in caudate volume showed the greatest effect size for both pre-HD (1·17) and manifest HD (2.04) groups compared to controls. Cognitive biomarkers all had relatively low effect sizes in pre-HD (<.20), but the symbol digit modality test (SDMT; Smith 1991), a measure of visuomotor integration with components of visual scanning and tracking, was the most sensitive cognitive measure in manifest HD (effect size = 1.00). Quantitative motor measures, designed to be objective quantitative measures of motor symptoms, were not as robust as imaging measures, with speeded-tapping mean inter-tap interval for the non-dominant hand having the largest effect size in pre-HD participants (.38), and the log of speeded-tapping tap duration variability for

non-dominant hand giving the largest effect size for manifest participants (.68). The clinical rating scale of motor progression showed larger effect sizes than quantitative motor measures. The UHDRS motor score had an effect size of .79 for pre-HD and .81 for manifest HD. Despite these and other results demonstrating that macrostructural MRI biomarkers are robust and sensitive measures of HD progression (Georgiou-Karistianis et al. 2013), the most widely used biomarkers are still clinical measures of progression that are more relevant to the daily life experiences of an HD patient, with MRI biomarkers currently being used as secondary endpoints for clinical trials (Scahill et al. 2012).

Recently, fluid biomarkers for HD have been gaining increased interest. While imaging and clinical biomarkers provide indirect measures of the disease, fluid biomarkers offer the potential to provide direct measures of biochemical changes occurring during the course of HD (Byrne & Wild 2016; Johnson et al. 2018). Biomarkers of cerebrospinal fluid (CSF) and plasma are currently being characterised, with neurofilament light protein (NfL), a biomarker thought to measure axonal degeneration, hypothesised to be a promising marker of cross-sectional disease stage but also a strong predictor of HD progression in the subsequent period (Byrne et al. 2017). NfL has also been shown to predict regionally specific atrophy in subcortical regions, regions of the cortex and the WM (Johnson et al. 2018). While showing great promise, fluid biomarkers require more validation, especially prior to use as surrogate disease markers in clinical trials.

1.1.2. Summary

HD is a devastating disease, characterised by a triad of slowly progressing motor, cognitive and psychiatric symptoms. There is no cure and no effective treatment, however clinical trials are underway aimed at developing disease-modifying treatments. At end-stage disease, atrophy in the caudate and putamen is the most pronounced neuropathological finding, however global neural atrophy suggests a slow progression of widespread neuronal damage. By using MRI to measure characteristics of the brain in-vivo, we can hope to better understand the trajectory of neural atrophy in HD.

1.2. MRI methods

MRI is a technique employed to examine the structure and function of the brain and body in-vivo. MRI scans allow measurement of neuropathological change as it occurs over repeated time points and without invasive procedures. While MRI scans are not used in the diagnosis of HD due to the availability of a genetic test the analysis of MRI data is regularly used in HD research studies and clinical trials to monitor the progression of neural change. An

MRI scanner is capable of collecting many types of data however this PhD will focus on sMRI scans, which provide a way of measuring structure and morphometry of the brain. A brief overview of MRI hardware and acquisition of images will be covered, followed by an examination of neuroimaging in HD.

1.2.1. Acquisition of MRI data

MRI scans utilise a magnet and radiofrequency pulses to change the state of hydrogen atoms in the body, and the energy created by these changes in state can be measured and outputted in the form of an image (Currie et al. 2013). Varying MRI acquisitions can be used to capture different characteristics and functions of the brain.

1.2.1.1. Hardware

An MRI scanner consists of two main components; the MRI scanner and the control console. The control console is the computer console that allows the programming and running of different scans. It is typically in a different room to the MRI scanner. The MRI scanner is a machine that uses a strong magnetic field that enables us to capture images.

The Maxwell equations stipulate that an electric current running through a wire creates a magnetic field surrounding the wire (Currie et al. 2013). An MRI scanner makes use of this principle, and uses a current running through a series of superconducting metal coils cooled to absolute zero to generate strong magnetic fields (called B0). The strength of the magnetic field is described in units of Tesla (T), with higher Tesla scanners usually resulting in higher resolution images. Most modern scanners are either 1.5T or 3T in strength.

A set of gradient coils are also present in the scanner, which are used to distort the main magnetic field in a predictable pattern and then allows spatial encoding of the measured signal so images can be mapped in the x, y and z directions to create 3D images. An MRI scanner also has radiofrequency (RF) coils that transmit an RF signal to the tissues being examined, and then receives the induced signal from the tissue. Finally, in order to collect good quality images, shim coils are used to ensure consistency, known as homogeneity, within the magnetic field. This translates to increased homogeneity within the MRI images when processed.

1.2.1.2. Obtaining images

MRI scanners use the hardware described to create images via the manipulation of atoms within the body. The source of most MRI images is the hydrogen nuclei. Within each

hydrogen nuclei there is a single proton with a positive electrical charge. Each proton is constantly spinning; since a spinning electrical charge is a current, each proton has a tiny magnetic field. These magnetic fields are called magnetic moments, and each magnetic moment is normally in a random orientation (Currie et al. 2013). However, when a uniform magnetic field (such as B0) is applied to magnetic moments they align to B0 in either a parallel or an antiparallel fashion. More protons will align parallel to B0 since this is the alignment requiring less energy, although the difference between the number of protons in parallel and antiparallel alignment depends on factors such as the strength of B0 and the temperature of the tissue.

In addition to aligning with B0, when an external magnetic field is applied to the protons they spin in a particular motion, called 'precession'. The protons precess at a frequency determined by the strength of B0 multiplied by γ, a constant value relative to the properties of each type of proton, this is called the Lamor Equation (Currie et al. 2013). The protons pointing in parallel and antiparallel directions cancel each other in all directions except for the z-axis of B0, so when a person is in B0 their hydrogen protons align with B0 creating a magnetic field for the patient aligned with B0. This field is called longitudinal magnetisation.

To create a signal a brief RF pulse is applied which creates a transfer of energy from the RF pulse to the protons, disturbing the protons so they fall out of line with B0 (Currie et al. 2013). To do this, the RF pulse must be applied at the same frequency as the frequency at which protons are precessing. This RF pulse reduces the longitudinal magnetisation and rotates the net magnetisation into the x-y plane as the protons absorb the energy, called excitation. It also causes the protons to precess in phase, i.e. in the same direction, at the same time. A simple schematic demonstrates this process in Figure 1.3.

Once the RF pulse has been switched off, the protons begin to return to their original alignment with B0 via two types of relaxation (Currie et al. 2013). The energy absorbed from the RF pulse is released and the protons move back into the z plane in line with B0. This process is called T1 relaxation. Each tissue has a different T1 relaxation time depending on how tightly bound the protons are in their environment. For example, within the CSF where hydrogen atoms and protons move more freely, the T1 relaxation time is longer. These tissue-dependent differences in relaxation time result in the ability to distinguish between different tissue types from MRI scans.

The protons also begin to fall out of phase once the RF pulse is switched off. This is called T2 relaxation, and is caused by the fact that the magnetic fields from neighbouring

protons influence each other and push each other out of phase. When combined with small inhomogeneity within B0, the result is differing T2 relaxation across the image.

T1 and T2 relaxation are independent, but occur at the same time. T1 relaxation happens more slowly than T2 relaxation, though. When the protons undergo T1 and T2 relaxation, they re-emit the energy absorbed from the RF pulse and this energy is measured by the MRI scanner and used to construct an image. The gradient coils described earlier are used to collect spatial information from the re-emitted energy. The raw data is captured from the scanner and stored in a data matrix called *k*-space. This data is then converted from *k*-space into an image via an inverse Fourier transform.

These two types of relaxation can be measured independently to create the most commonly used types of structural images, T1 and T2 weighted images. T1 images measure the T1 relaxation and T2 images measure the T2 relaxation and have different contrast based on the properties of the image acquisition. Figure 1.4 shows an example of a T1 and T2 image from the TRACK-HD study, demonstrating the different contrast seen in these two scan types. An MRI acquisition can specify not only the contrast in the image, but the quality of the image, and the size of the voxels. Image quality and MRI artefacts are discussed further in Chapter 2.

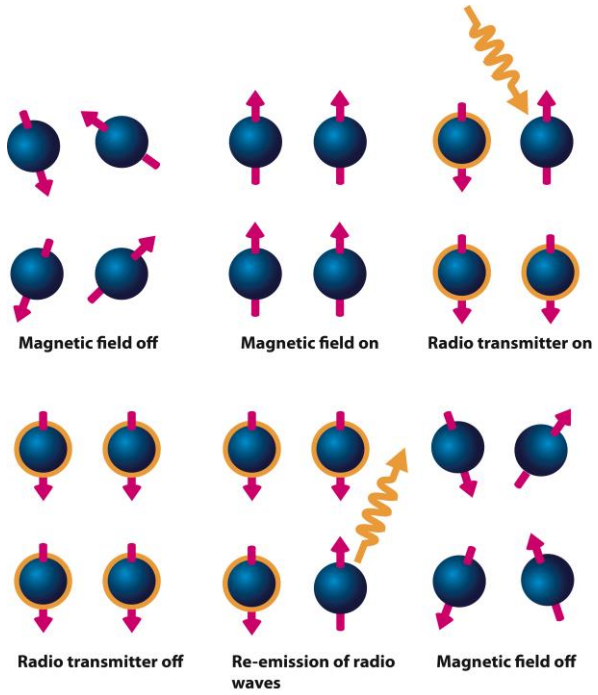


Figure 1.3. A schematic illustrating the spin of hydrogen atoms before a magnetic field is applied, whilst in a magnetic field, and during and after an RF pulse is applied. Adapted from https://www.khanacademy.org/test-prep/mcat/physical-processes/proton-nuclear-magnetic-resonance/a/magnetic-resonance-imaging-mri.

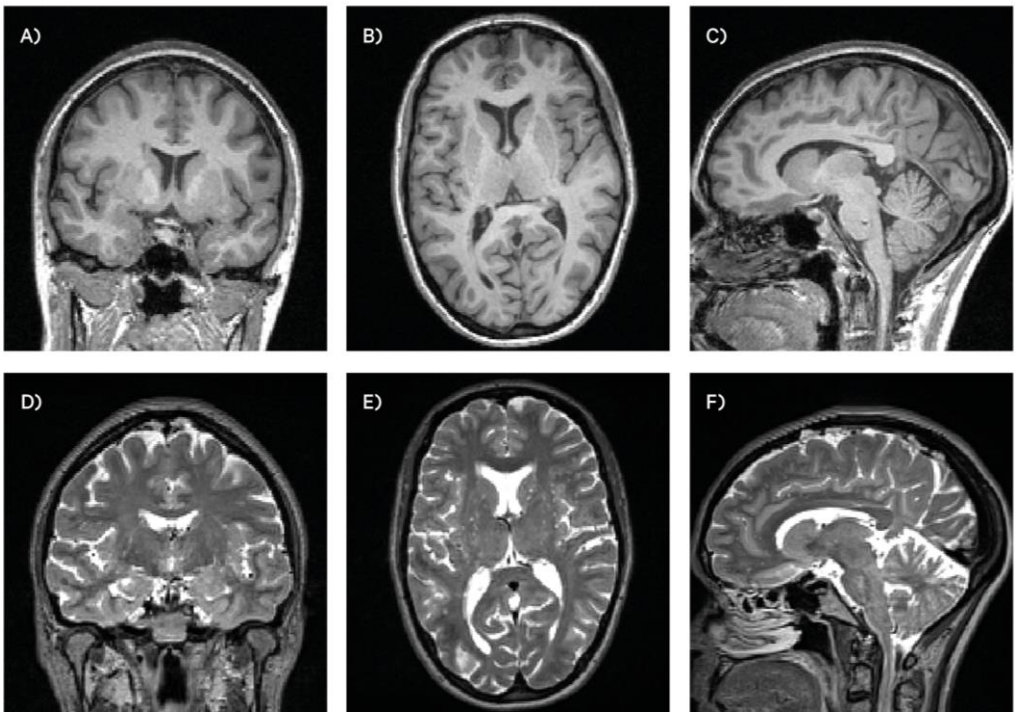


Figure 1.4. An example of 3D T1 and T2 weighted images from the TrackOn-HD study showing (A) A T1 scan in coronal view, (B) A T1 scan in axial view and (C) A T1 scan in sagittal view, (D) A T2 scan in coronal view, (E) A T2 scan in axial view and (F) A T2 scan in sagittal view.

1.2.1. Types of MRI data

A number of types of MRI data can be collected from MRI scanners, with each type representing different neural characteristics. Three types of data frequently collected for use in HD research are structural, diffusion and functional MRI data.

1.2.1.1. Structural

sMRI provide static anatomical information about the brain. They are typically used to measure characteristics of different brain regions, such as volume, both cross-sectionally and longitudinally and will be the main focus of this research. sMRI is currently the most robust type of MRI imaging to be replicated reliably across multiple study sites (Georgiou-Karistianis et al. 2013), and structural imaging measures show the greatest effect sizes in detecting longitudinal HD pathology (Hobbs et al. 2013). The majority of studies that have previously reported on sMRI changes within HD participants have used three main techniques: regional volumetric, voxel-based morphometry (VBM) and CT analyses, to be discussed in section 1.3.

1.2.1.2. Diffusion

Diffusion tensor imaging (DTI) detects the coherence of water flow through tracts of the brain, and is hypothesised to reflect aspects of tissue microstructure such as myelination

(Le Bihan et al. 2001). A number of different metrics can be extracted from DTI images representing different properties of the tracts such as fractional anisotropy, which represents overall coherence, and mean diffusivity, which represents mean diffusion along the tracts. In HD individuals, changes in these characteristics are believed to represent underlying pathology that precedes change detected in the WM by sMRI.

1.2.1.3. Functional

Functional MRI (fMRI) quantifies increases or decreases in blood flow (Blood Oxygenation Level Dependent signal) to regions of the brain during rest or the performance of different tasks (Ogawa et al. 1990). The regional changes in blood flow are due to the varying demand for oxygen across different brain regions. Differences between groups or over time are often attributed to differences in pathology. It is hypothesised that as HD progresses, alternative brain networks compensate for damaged networks to complete a task or function thus resulting in differences in BOLD signal between controls and HD gene carriers (Klöppel et al. 2015; Gregory et al. 2017). fMRI can be collected in a number of ways, with two commonly used paradigms. Task based fMRI measures BOLD signal during the completion of a task in the MRI scanner, for example, increases or decreases in signal could be measured during a finger tapping task. Resting state fMRI (rs-fMRI), in contrast, measures BOLD signal during rest in the scanner, and is rs-fMRI data quantifies the underlying blood flow and regional interactions taking place at rest.

1.3. Analysis of sMRI data

The processing of sMRI data to quantify group differences or longitudinal change can be divided into two broad approaches, region of interest (ROI) analyses and whole-brain analyses. Both methods offer complementary information, and within these broad techniques there are a multitude of different measurements that can be performed.

1.3.1. Region of interest analysis

ROI analyses involves the delineation of structural regions, for example the whole-brain or caudate, to provide precise measurements of differing characteristics of a region. Measures such as volume, thickness of the GM cortex or gyrification can all be examined using ROI studies. The delineation of ROIs can be performed manually or automatically (see section 2.3 for an explanation of different processing methods), and after a region is delineated a value can be extracted. Figure 1.5A shows an example of a manually delineated caudate ROI used to measure caudate volume. ROI studies often show high sensitivity for detecting group

differences and longitudinal change and thus can be very useful for examining specific hypotheses, however, in order to obtain accurate data the ROIs need to be precisely defined. The manual measurement of ROIs is a highly accurate method for measuring the characteristics of some regions, yet they are also very time-consuming and can be impractical when measuring multiple anatomically complex regions. Automated measures can be performed more quickly and easily, however they often bring an increased level of error. These costs and benefits should be weighed up for each study using ROI methods.

In addition to issues with measurement, the nature of ROI studies can result in group differences or volumetric change remaining unobserved due to the highly specific nature of the analysis. In order to avoid harsh statistical corrections for multiple comparisons, regions should be carefully selected for examination prior to analysis based on *a priori* hypotheses. However, this can result in significant results being overlooked if they occur outside of the chosen ROIs. ROI examinations are frequently used in HD studies, especially when measuring subcortical structures such as the caudate and putamen. A number of HD studies use manual ROIs to characterise change (Tabrizi et al. 2009; Tabrizi et al. 2012; Tabrizi et al. 2013; Tabrizi et al. 2011), and some of these measures are described in 2.3.3. Often, additional exploratory whole-brain methods are used to supplement *a priori* ROI investigations of sMRI data.

1.3.2. Whole-brain analysis

Whole-brain techniques are used to perform comparisons between groups or longitudinally across all regions of the brain, Figure 1.5B shows the results of a whole-brain group analysis using VBM (section 2.3.4.2.1). Whole-brain analyses involve the statistical comparison of every voxel across the brain, and thus the performance of mass-univariate statistical tests require strict correction for multiple-comparisons (see section 2.4.4). However, whole-brain analyses offer the chance to look for regional differences that may be overlooked in ROI analyses. They have been widely used in HD imaging studies to measure both volume and cortical thickness (CT), and help to provide an understanding of the different regions that may be affected by atrophy in HD. Whole-brain analyses generally offer limited flexibility in terms of design (e.g. the options can be limited when using data sets with multiple time points, or missing data) and thus can be limited in their applications.

VBM is the most commonly used whole-brain approach in HD, with SPM the most commonly used software. This involves a series of steps (Ashburner & Friston 2000), with the first step an integrated process registering all scans to create a template whilst also segmenting them into three tissue types (GM, WM and CSF). Following this, the scans are

aligned to the template space created in the first step, all voxels are modulated based on the warping parameters to account for volumetric changes that may have occurred during registration, and smoothed to reduce errors associated with registration. A general linear model statistical comparison can then be specified that can be used to compare voxel-wise volumetric differences between participants, or to examine associations between volume and performance on a cognitive task.

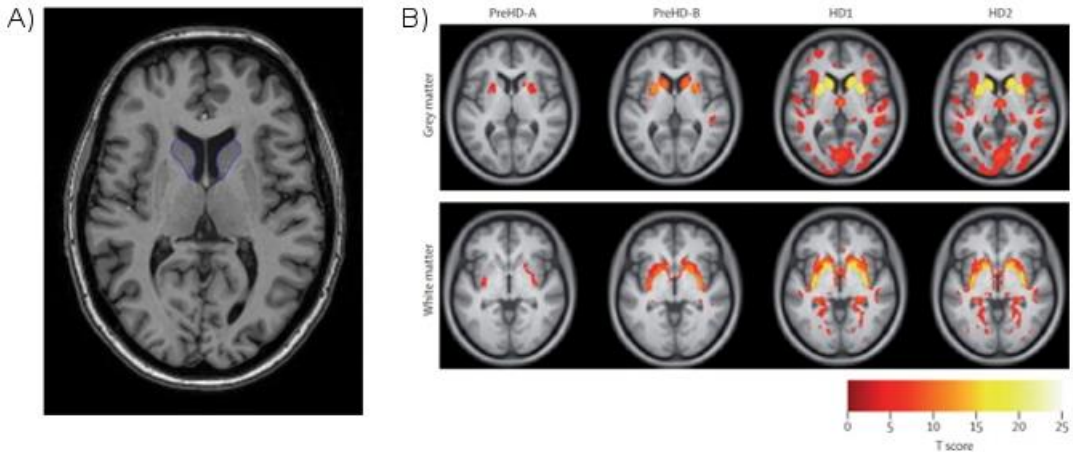


Figure 1.5. (A) Shows a manually delineated caudate region on a scan from the TRACK-HD study; (B) Shows the results of a cross-sectional whole-brain analysis comparing the volume of GM and WM in different HD groups to controls from the TRACK-HD study (Tabrizi et al. 2009).

1.4. sMRI findings in HD

Both cross-sectional and longitudinal sMRI studies aimed at characterising neural atrophy in HD have been performed on a range of cohorts. Generally, the results suggest that prior to diagnosis there is significant subcortical GM loss and some WM loss (Tabrizi et al. 2009; Paulsen et al. 2008), but that cortical atrophy begins closer to the time of diagnosis (Tabrizi et al. 2009). Atrophy across all tissue types and regions continues to advance after diagnosis and throughout the course of the disease (Tabrizi et al. 2013; Tabrizi et al. 2011; Tabrizi et al. 2012; Aylward et al. 1997). These findings will be covered in more detail here.

1.4.1. Regional atrophy in HD

A large amount of research has been focused on characterising atrophy in basal ganglia regions based on early pathology studies showing that the basal ganglia undergo particularly pronounced atrophy in HD (Vonsattel et al. 1985). Findings from MRI studies indicate that the striatum is the earliest region to undergo atrophy (Nopoulos et al. 2010; Aylward et al. 1997; Aylward et al. 2004; Tabrizi et al. 2009; Tabrizi et al. 2011). Cross-sectional

differences between controls and HD gene carriers can be found in the striatum up to 15 years prior to disease onset (Paulsen et al. 2010), and work is currently underway to examine participants who are even further from onset to determine when differences in the striatum can be detected. Putamen volume is also significantly lower in pre-HD participants than in control participants more than 15 years prior to symptom onset (Majid et al. 2011; Paulsen et al. 2010). There is currently a lack of clarity as to whether the caudate and putamen begin to undergo atrophy at the same stage and rate, or whether they progress differently, with different studies showing slightly different patterns of atrophy. The divergence between results is probably due to differing cohorts and measurement techniques (Georgiou-Karistianis et al. 2013). Despite this, it is clear that for both putamen and caudate structural differences continue to get more pronounced with increasing disease progression.

Additionally, longitudinal studies have found significant striatal volume change in HD over very short periods of time. While annual rates of change can vary between studies depending on the cohort and analysis method, rates of around 3-4% reduction per year have been reported (Georgiou-Karistianis et al. 2013). Furthermore, volumetric change in both the caudate and putamen is significant over 12 months in both pre-HD and manifest HD compared to controls (Tabrizi et al. 2011; Hobbs et al. 2015), and over 6 and 9 months in the caudate in manifest HD compared to controls (Hobbs et al. 2015).

Beyond the striatum, differences in whole-brain volume can also be detected between controls and pre-HD participants who are less than 10 years from disease onset (Tabrizi et al. 2009). Additionally, whole-brain volume change is significant over 1 year in these participants, and significant over 6 months in early HD patients (Henley et al. 2006; Tabrizi et al. 2011; Hobbs et al. 2015). Change in total WM is also detectable over one year in participants more than 10 years from disease onset, with the rate of atrophy increasing closer to and after disease onset (Tabrizi et al. 2011). In manifest HD, WM volume shows significantly greater decline over 6, 9 and 15 months than in control participants (Hobbs et al. 2015). Small differences in total GM volume (with total GM including both subcortical and cortical GM) have been found between controls and pre-HD participants who are over 10 years from disease onset, however the greatest GM changes occur in participants close to disease onset and those with manifest HD (Nopoulos et al. 2010; Tabrizi et al. 2012), with total GM again showing significantly greater change in as little as six months in manifest HD compared to controls (Hobbs et al. 2015).

In addition to quantifying change over time, many of the volumetric findings have been tested as predictors of disease progression. In the TRACK-HD cohort, striatal and GM

volume measures were able to predict diagnosis in pre-HD individuals, and whole-brain, caudate, putamen, ventricular and GM volumes can significantly predict decline in total functional capacity in early HD participants, supporting a strong relationship between neural atrophy and clinical change (Tabrizi et al. 2013). Furthermore, in the PREDICT-HD cohort, cross-sectional and longitudinal volume of the putamen, hippocampus, CSF, accumbens, globus pallidus and caudate could predict motor onset (Paulsen, Long, Ross, et al. 2014). Figure 1.6 shows a schematic diagram of the hypothesised progression of imaging biomarkers in relation to CAP score (Ross et al. 2014).

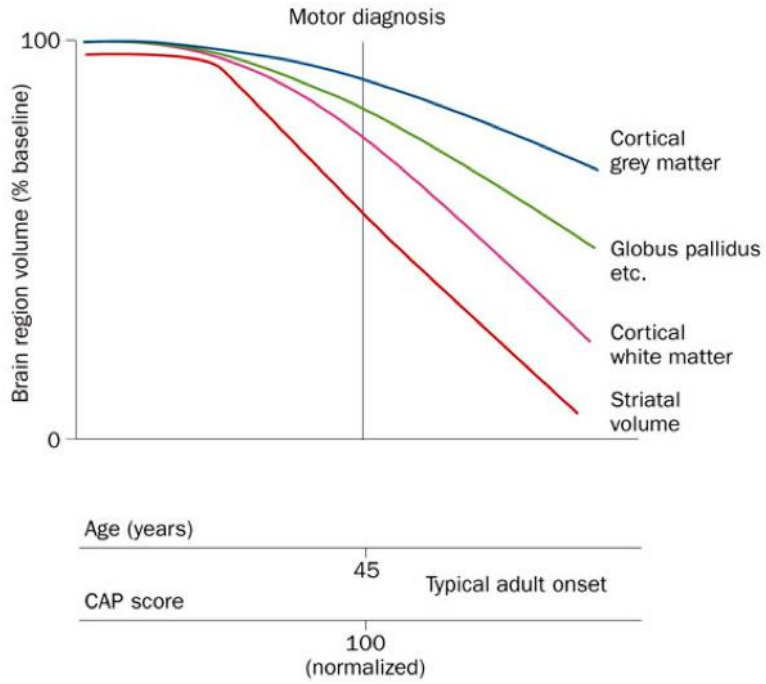


Figure 1.6. A schematic representing hypothesized progression of imaging biomarkers commonly used in HD. Adapted from Ross et al. (2014).

1.4.2. Cortical atrophy in Huntington's disease

As previously described, widespread cortical atrophy is regularly reported in post-mortem studies of end-stage HD patients. However, in-vivo imaging studies have found inconsistent results when measuring cortical change. While results suggest that atrophy is occurring by early-stage HD, the quantification of cortical atrophy in different cohorts via different techniques has resulted in a range of findings.

Most studies examining atrophy in the cortex have been cross-sectional whole-brain studies performed using VBM. Within the TRACK-HD study, reduced cortical volume has been detected in the cingulate, precentral and prefrontal regions as well as the occipital, parietal,

and small regions of the temporal lobes in manifest HD compared to controls (Tabrizi et al. 2009). In addition to the findings shown in the TRACK-HD cohort, group differences between pre-HD and control participants have been described in other cohorts within the motor regions, as well as within occipital and frontal regions (Thieben et al. 2002; Gómez-Ansón et al. 2009; Sormani et al. 2004). Between controls and manifest HD participants, there have also been significant differences reported in the precentral and postcentral regions (Douaud et al. 2006; Kassubek et al. 2004; Wolf et al. 2009), as well as in frontal and parietal regions (Wolf et al. 2009; Gavazzi et al. 2007). Atrophy of the frontal lobe has also been reported in an early ROI study, with a volume reduction of 17% in early-HD participants compared to controls (Aylward et al. 1998).

Longitudinally, both pre-HD participants nearing onset and manifest participants from the TRACK-HD cohort showed significantly greater cortical change than control participants over 24 months. Volumetric reduction in the occipital cortex was the most striking change over 24 months in pre-HD and manifest groups, with additional atrophy occurring in parietal and frontal regions in manifest participants (Tabrizi et al. 2012). One other longitudinal study reported increased atrophy in the occipital, frontal and parietal cortices over 24 months in early-HD participants compared to controls (Hobbs, Henley, et al. 2010). In addition, another VBM study detected atrophy over 12 months in the caudate, pallidum, putamen, insula, cingulate cortex, cerebellum, orbitofrontal cortex, medial temporal lobes and middle frontal gyri. However, this was in a cohort of 49 participants that included juvenile HD participants (as young as 7 years old) who show a different pattern of symptom progression, with juvenile HD often undergoing more rapid disease acceleration then adult onset HD. It is likely that registering and analysing juvenile and adult HD brain scans together would result in differing findings than when just measuring change in adult onset HD, and thus the results are not representative of a true pattern of adult onset HD progression (Ruocco et al. 2008).

To attempt to collate and summarise the results of 17 VBM studies performed on HD cohorts a meta-analysis was conducted in 2013. After the results from all studies were combined, the only cortical region to show differences between pre-HD and controls was a region in the right occipital lobe (Dogan et al. 2013). In manifest HD, there were significant differences with controls found in the frontal cortex, primary motor, premotor and somatosensory regions, as well as in the intraparietal sulcus, the midcingulate cortex and the secondary somatosensory cortex (Dogan et al. 2013). The results from this meta-analysis suggest a pattern of atrophy that begins prior to onset in the occipital lobe, and extends to the frontal lobe, key motor regions and further into the occipital lobe after onset of HD. Taken

together, the results of these studies provide evidence that atrophy is occurring in both pre-HD and manifest HD participants, with frontal, motor regions and occipital regions showing significant group-differences or change in a number of studies. However, they do not provide detailed information about the timing or regional progression of atrophy.

As well as volumetric change, a number of studies have measured CT in pre- and manifest HD compared to controls. CT is the distance between the WM/GM boundary and the GM/CSF boundary. Significant differences in the thickness of GM have been reported between controls and HD gene carriers in both pre-HD and manifest stages of the disease (Rosas et al. 2008; Rosas et al. 2002; Rosas et al. 2011). Results suggest that the earliest change in CT is in the occipital and posterior regions of the brain, with more anterior regions affected later in the disease (Rosas et al. 2002; Rosas et al. 2011). However, CT studies are almost exclusively cross sectional. One longitudinal study of CT found significant change in early HD participants over one year, with the main changes seen in sensorimotor and fronto-parietal motor regions. However, the number of participants in this study was small (N=22, split into three groups), there was no control group, and the authors did not control for multiple comparisons (Rosas et al. 2011). A more recent study using a large cohort to study the effect sizes of different imaging biomarkers, reports poor longitudinal sensitivity for CT in HD, especially when compared to volumetric measures, indicating that perhaps CT is not a sensitive measure of neural change in HD (Hobbs et al. 2015).

Recently, more complex methods of characterising GM change have been applied to HD cohorts with the aim of better understanding the progression of atrophy in HD. Two studies have attempted to map the co-variance of GM changes in HD. Co-variation analyses are designed to estimate the dependencies between regions of the brain showing volumetric reduction in HD, which are then thought to represent networks that undergo atrophy simultaneously. One study conducted a large-scale analysis of pre-HD participants (N=831) compared to controls (N=219), initially performing VBM in SPM to determine regions of lower GM concentration, and then performing source-based morphometry (SBM; Ciarochi et al. 2016). SBM uses a whole-brain multivariate approach to capture co-occurring patterns of different GM concentration across the brain in HD compared to controls. SBM analysis involves performing the same processing steps as VBM, followed by an independent component analysis (ICA) to identify regions of change that occur together, called components (Xu et al. 2009). An ICA is a statistical approach used to categorise sub-components of a measure into larger hidden components that are thought to represent over-arching constructs. Statistical analyses can then be used to determine which components show a significant pattern of co-

occurrence and to test relationships between these components and other variables (Xu et al. 2009). From the VBM analysis, the authors report significant group differences in GM concentration in the caudate, frontal lobe, occipital lobe, as well as the hippocampus and thalamus. A number of components were identified by the SBM, and these regions were then related to HD disease stage. Participants were divided into groups based on estimated time until onset, with far, medium and close to onset groups. One of the components identified by the SBM was made up of frontal and some motor regions (precentral, supplementary motor), and this was the earliest component showing atrophy in the far from onset group. In addition, parietal and occipital atrophy was found to be co-occurring, but in those closer to onset. A range of other components showed increasing atrophy with proximity to onset. While these results indicate interesting patterns of co-occurring GM loss, visual inspection of the VBM results suggests a spatial misalignment during processing as the VBM map is poorly aligned with the template. This could mean that the results from both the VBM and SBM have been mislabelled and thus should be interpreted with caution. In addition, the authors reported GM concentration on unmodulated data. Modulation is a processing step that aims to preserve differences in morphology which may have been affected by the inter-subject alignment performed during processing of VBM data (Henley et al. 2010; Keller et al. 2004). By failing to modulate the data these results are likely to be less sensitive to neuroanatomical differences than in data that is modulated. Together, these issues mean that the results of both the VBM and the SBM analysis could be inaccurate, and thus require further validation.

A second study with 30 pre-HD participants, 30 HD participants and 30 controls also looked at structural co-variance of GM changes, using a similar methodology but via a technique included as part of FSL's software library (Coppen et al. 2016). This study also performed VBM first, with the greatest differences between pre-HD and controls seen in the caudate and putamen and small differences found in the cortex in the insula and in the parietal regions of the planum temporale, parietal operculum as well as the posterior supramarginal gyrus. Cortical differences were more widespread when comparing the controls to HD participants, with significant differences found in the precentral and postcentral regions, supplementary motor cortex, lateral occipital cortex and the frontal pole. The results of the SBM identified 10 structural GM networks, with two of these networks showing significantly reduced network integrity in both pre-HD and HD compared to controls. One network incorporated the caudate, the nucleus accumbens, putamen, pallidum and precuneus. The other included the parahippocampul gyrus, the cerebellum, the pallidum and the planum polare. Furthermore, one network made up of the precuneus, cuneus, lateral occipital and

lingual regions showed significantly lower network integrity in HD compared to controls. The reduced structural integrity seen in these three networks was interpreted by the authors as a demonstration of the general breakdown of associated GM structures in HD. Again, these results demonstrate an interesting pattern of co-variation in GM change, indicating early change in parietal, motor and occipital regions, however the linear modelling used in this analysis does not allow for possible accelerations or decelerations in change across the disease. In addition, by measuring the co-occurrence of regional differences, it is assumed that that within each network the regions undergo the same temporal progression within the same disease stage. However, it is biologically plausible that different regions show differing rates of atrophy and variable patterns of acceleration and deceleration. Until there is a detailed understanding of the overall pattern of atrophy, the measurement of network changes could also be seen as premature since more basic information about the progression of cortical atrophy is not yet well understood.

Another recent study used a different methodology to examine the pattern of brain changes via both sMRI and diffusion MRI on data from the PREDICT-HD cohort. The authors used a multivariate linear regression model to detect inflections points (change-points) in different measures in relation to predicted years to onset. The change-points represent changes in the linear trajectory of a measure, reflecting the disease stage at which each measure begins to show significant change in pre-HD. The data was a cross-sectional selection of 85 controls and 212 pre-HD participants for the sMRI analysis, and 79 controls and 178 pre-HD participants for the DTI measures. Significant volumetric loss and significant change points were seen in both subcortical and WM regions, suggesting that subcortical and WM regions undergo significant volumetric decline in different stages of pre-HD. No cortical regions had significant change-points. However, significant volumetric expansion and change-points were seen in a number of sulcal regions, including the occipital and Sylvain fissures. The authors argued that sulcal expansion is an indirect measure of cortical change. This study used a novel framework to compare different neuronal markers of disease progression, and while the integrated analysis of subcortical, WM and cortical measures as well as diffusion metrics is required to understand the relationship between different indicators of neural change, it is possible that the large number of regions and contrasts resulted in poor power to detect cortical change, especially based on previous findings suggesting that cortical change is likely to be more subtle than that seen in other regions (Tabrizi et al. 2012). It is important that more work is done to characterise the overall pattern of CGM change prior to integrating it with other metrics of neural change.

1.4.3. Limitations of these findings

Despite the amount of research that has been designed with the aim of characterising neural change in the cortex in HD, the conclusions we can draw from these studies are limited by several factors. As described, there are a number of different software packages available to calculate sMRI measures. While several studies have previously examined methods of quantifying these characteristics, it is unclear which CGM characteristics and which software packages provide the most accurate and sensitive measures of between-groups differences and within-groups change in CGM in HD. The majority of available software used to characterise the CGM was developed and validated on healthy participants (Irimia et al. 2012). Applying these tools to brains showing neural pathology, such as the scans from HD participants, can result in poor performance and inaccurate measures of CGM characteristics (Irimia et al. 2012). Recent work has identified poor segmentations as a possible driver in variability within volumetric sMRI studies across neuroimaging (Ashburner et al. 2016). Furthermore, a number of tools commonly used in HD to quantify cortical atrophy show issues with accuracy (Katuwal et al. 2016). It is thus vital that prior to measuring CGM atrophy in HD the tools used to calculate CGM characteristics are evaluated.

Additionally, most previous work has utilised simple univariate analyses calculating VBM or CT measures. These investigations do not allow for complex analysis of several time points, or for the examination of non-linear patterns of accelerations or decelerations in atrophy rates. They can also suffer from potential over- or under-correction for multiple comparisons. Multivariate analyses provide a more powerful approach to infer network-wide patterns of atrophy and because they are not limited by the use of stringent corrections for multiple comparisons (McIntosh & Misic 2013; Habeck 2010). However, to date, the multivariate analyses performed examining CGM change have all used cross-sectional data and models that make an assumption of linearity in the progression of atrophy. In addition, one of the studies has used data from different scanner strengths across >20 sites, with another using from 23 sites. While the use of multiple sites is common in neuroimaging studies, there is a risk that subtle cortical atrophy may not be detected due to between-scanner variance and noise.

A final issue that limits the conclusions from these studies concerns the division of participants into groups based on estimated years to onset. While group separation in HD studies is commonly done via calculated estimates of predicted years to onset, different algorithms and variable definitions of groups are used between studies. Furthermore, since predicted years to onset is not always an accurate measure of onset, heterogeneity within groups could be masking the earliest atrophy in the cortex. Most previous studies in HD have

tended to maintain groups based on status at study baseline rather than re-categorising or separately examining those participants who undergo diagnosis during the study. A small number of studies have observed brain changes over transition to HD (Tabrizi et al. 2013; Aylward et al. 2012), with one study investigating global brain change and the other striatal change. Since previous results suggest that there is an increase in CGM atrophy around the time that symptom severity also increases (Tabrizi et al. 2012), it appears that CGM atrophy could be a factor in the progression of HD symptom onset. CGM atrophy has implications for biomarker development, the timing of interventions and even the identification of new therapeutic targets, thus it is important that a thorough examination of participants undergoing the transition from pre-HD HD to manifest HD is conducted.

1.5. Theories of the progression of neural pathology in HD

There are a number of theories of how HD pathology develops throughout the brain. These have been generated from current understandings of HD biology, as well as from imaging findings in HD and other neurodegenerative diseases. Most theories first depend upon an understanding of basal ganglia cortical networks, described briefly here.

The basal ganglia includes the striatum (caudate and putamen), substantia nigra, the subthamanic nucleus and the globus pallidus (Jahanshahi et al. 2015). The basal ganglia was initially thought to be involved only in motor tasks, but is now understood to also contribute during both cognitive and emotional processing, such as during reinforcement learning, decision making and a number of other tasks. There are three main circuits that are thought to be driving performance on these domains, the motor, associative (cognitive) and limbic (emotional) circuits (Jahanshahi et al. 2015). These three networks utilise the basal ganglia and different cortical regions, and are often required to integrate in order to successfully perform actions. Figure 1.7 shows a schematic of these three networks.

In HD, a number of theories have been hypothesised that link GM, WM and functional changes to a disruption within these circuits. One theory that was proposed for multiple neurodegenerative diseases suggested that pathology might spread based on a prion hypothesis of disease progression. The hypothesis of prion-like disease progression argues that pathology spreads throughout the brain in a cell-to-cell manner, and thus brain regions that are close to the epicentre of disease or that are strongly linked to that area will undergo degeneration prior to those further away or less closely linked. While a number of researchers have argued that their work demonstrates evidence of this theory (Raj et al. 2012; Raj et al. 2015; Zhou et al. 2012) because they have shown that regions in close proximity to the disease

epicentre or highly connected regions undergo damage earlier than unconnected regions, this theory has not been directly demonstrated, especially in HD.

Alternatively, McColgan et al. (2017; 2015) proposed that instead of prion-like spread, the progression of atrophy in HD could be related to connectivity across the brain, with regions that have more connections to other brain regions (rich club regions) undergo greater damage due to the metabolic demand placed on these regions and connections. McColgan (2015) found that cortico-striatal regions were more affected in those with HD, showing reduced connectivity between striatal regions and both frontal and parietal/occipital regions. These regions are rich club regions, and also part of the motor and associative circuits, and could indicate a susceptibility for degeneration in HD within these two networks that are frequently recruited in day-to-day functioning.

These theories do not explain the degeneration commonly reported within other regions in HD, especially within the occipital cortex. There is no hypothesis that has provided sufficient evidence to explain degeneration in this region, especially given that visual disturbances are not typically reported in HD. The most commonly cited suggestion is that atrophy within this region is due to high metabolic demand placed on this region (Rosas et al. 2008; Feigin et al. 2001). While there is no conclusive evidence to explain the pattern of pathological spread in HD, as imaging methodology advances it is hoped that new techniques will continue to improve our understanding and to support work investigating the biological components of HD.

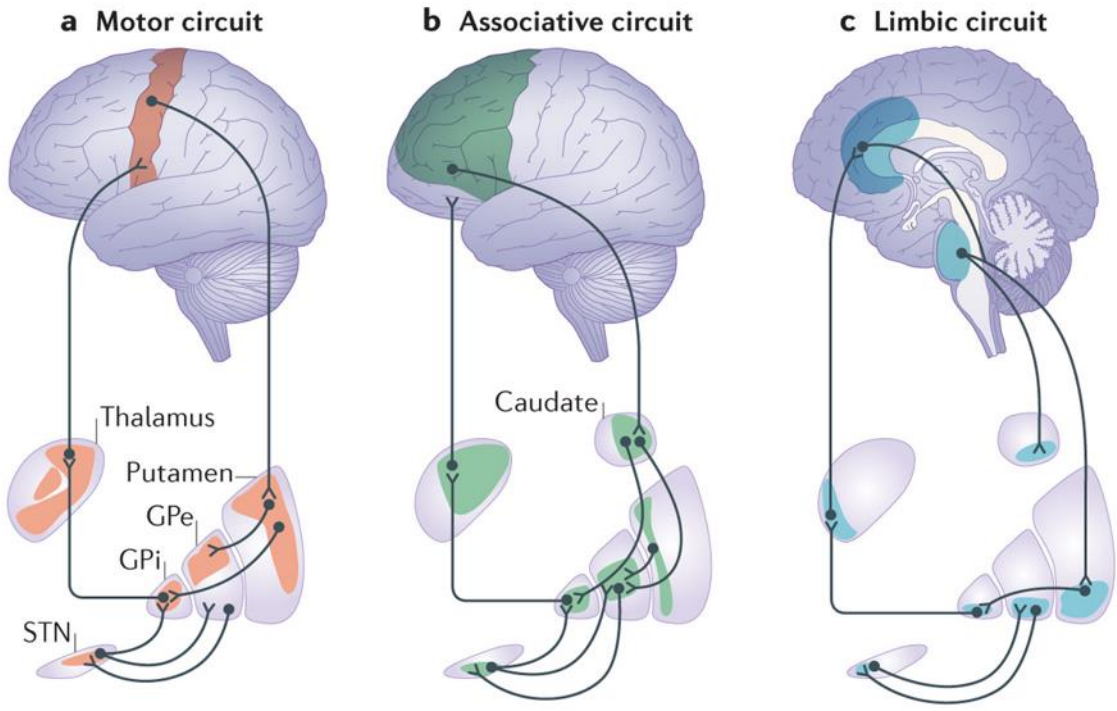


Figure 1.7. A schematic demonstrating the three proposed circuits connecting basal ganglia to the cortex (Jahanshahi et al. 2015).

1.6. The current thesis

The aims of this thesis have been motivated by the lack of research that has performed detailed characterisations of the cortical changes in the period surrounding HD motor diagnosis. While a range of work has studied atrophy in pre- and manifest HD, most analyses have utilised univariate techniques applied on cross-sectional cohorts, offering limited conclusions. More recently there has been a shift towards multivariate techniques, albeit mostly still using linear models, however the range of participants included in these studies and the use of a predicted onset could mask the detection of cortical atrophy. Furthermore, the techniques commonly used to quantify CGM in these studies require detailed validation in HD. This work aims to undertake a comparison of CGM methods and to then provide a detailed analysis of CGM changes in HD motor converters.

First, a thorough validation of the available tools for CGM measurement will be undertaken. Using the results of this validation, the most sensitive and accurate methods for quantifying disease-related CGM change in HD will be selected. These methods will then be applied in an examination of the distribution and trajectory of CGM atrophy during motor onset in HD. The investigation of cortical atrophy will be performed on a sub-set of pre-HD participants who underwent conversion to manifest-HD during data collection for the TRACK-

HD and TrackOn-HD studies. The analysis will use a dynamical approach to perform structural modelling. This newly developed technique enables the quantification of net atrophy in a given time period, the rate of linear atrophy as well as non-linear accelerations or decelerations of atrophy. By understanding the progression of transition phase cortical change, we gain a greater understanding of the progression of HD more generally. The results will provide novel and clinically relevant information helping to characterise and understand the process of neural atrophy in HD.

2. GENERAL METHODOLOGY

This chapter will describe the general methodology used throughout this thesis. Further elaboration will be provided within each chapter for the specific application of these methods. The cohorts used in the analyses will be covered, along with information on the MRI acquisition parameters, pre-processing steps and methods of quantifying outcomes from sMRI data. Finally, some of the statistical methods used in this thesis will be described.

2.1. Cohorts

MRI data from four sources were used during this thesis. Freely available phantom data from the BrainWeb dataset were used for the methodological comparisons conducted in Chapter 3. Data from the TRACK-HD, TrackOn-HD and PADDINGTON studies were also used throughout this thesis. Descriptions of these cohorts and the data are provided here.

2.1.1. BrainWeb

Due to the nature of in vivo MRI scans it is impossible to measure the ground truth (GT) of various brain characteristics, such as the 'true' value of GM volume. Even established manual brain measures are likely to include some degree of error. Because of this, a number of the comparisons included in the methodological development section of this thesis were initially run on simulated data from the BrainWeb dataset (http://www.bic.mni.mcgill.ca/brainweb/; Aubert-Broche et al. 2006). The simulated data provides artificially created MRI scans and corresponding GT segmentations. The GT segmentations are regions used to create the phantom MRI scans and are designed to enable comparison with the performance of different software tools. The BrainWeb phantom data can be processed by different software, and the results can be compared to the GT regions to determine the accuracy of the tools. The BrainWeb data is artificial and based on healthy brains and so is not a true representation of human MRI data, especially in brains with abnormal pathology, yet it is a useful way of examining software performance. The BrainWeb dataset used in the current study contains a set of 20 freely available simulated brains (Aubert-Broche et al. 2006). The participants were 10 females and 10 males, with a mean age of 29.6 years and an age range of 24-37 years. Construction of the data is detailed in section 2.2.1.1.

2.1.2. TRACK-HD

TRACK-HD was a longitudinal multi-site observational study that began in 2008. Participants returned on a yearly basis, with the concluding visit in 2011. TRACK-HD aimed to recruit 360 participants at baseline, 120 control participants, 120 pre-HD participants, who had tested positive for the genetic expansion that causes HD but were not yet symptomatic, and 120 participants with manifest HD.

Recruitment of people at risk from HD was limited to those who had previously undergone genetic testing, with preliminary screening involving the self-report of genetic result. At baseline of TRACK-HD, gene positive participants were required to have a positive genetic test of ≥40 CAG repeats, which was confirmed via a blood test conducted at their first visit.

Pre-HD and HD participants were also required to have a burden of pathology score > 250 (as described in section 1.1.4.1; Penney et al. 1997). The burden of pathology score uses two of the most significant predictors of disease onset, age and CAG repeat, to estimate the burden of disease. A higher score indicates a higher burden and represents an individual's lifetime exposure to MTT. This criterion ensured that pre-HD participants were not too far from predicted disease onset to detect meaningful group differences when compared to controls; including participants who were further away from disease onset would have required larger group sizes to detect differences. Pre-HD participants were also required to have a Total Motor Score (TMS) on the Unified Huntington's Disease Rating Scale (UHDRS) of < 5 (Huntington Study Group 1996). The TMS measures the presence of motor symptoms with a score of <5 indicating no substantial motor symptoms.

The pre-HD cohort was separated into two groups based on the median expected years to disease onset as calculated by a survival analysis formula previously described (Langbehn et al. 2010); those estimated to be more than 10.8 years from disease onset were classified as the preHD-A and those less than 10.8 years from estimated onset preHD-B. Using the UHDRS (Huntington Study Group 1996) the HD cohort was also split into two groups. The participants were classified based on their Total Functional Capacity (TFC) scores as stage 1 (HD1:TFC=11-13) or stage 2 (HD2:TFC=7-10). TFC measures someone's ability to perform everyday activities such as domestic tasks, handling finances and working. The HD1 group included participants who showed higher functioning and were thus less affected by HD, and the HD2 group consisted of participants who were more affected by HD and had lower daily functioning scores. Since TMS <5 for this study, no participants showed substantial motor symptoms at baseline. Education was measured using the International Standard Classification

of Education (ISCED; UNESCO 1997). The ISCED comprises seven levels of education ranging from 0 ("pre-primary education") to 6 ("second stage of tertiary education, leading to an advanced research qualification").

Participants in the control group were selected to maintain as much consistency as possible with the gene carriers in terms of shared environment and thus the group was comprised of age- and sex-matched partners, spouses and gene-negative siblings of the gene-carriers. In addition to the criteria described, participants were required to be aged between 18-65, be able to tolerate both MRI and biosample collections and be suffering from no major psychiatric disorder or have a history of significant head injury at the time of recruitment. Generally, medication was not an exclusion factor unless it was medication being taken as part of a therapeutic trial. Other comorbid medical conditions were also generally accepted, unless they were deemed to impact upon a participant's ability to take part in the study. Ethical approval was given by the local ethical committees and written informed consent was obtained from each participant according to the Declaration of Helsinki. Full selection criteria and data collection processes have been published previously (Tabrizi et al. 2009).

At the end of the baseline visit, the TRACK-HD study had successfully enrolled 366 participants, made up of 123 controls, 120 pre-HD and 123 HD participants.

Table 2.1 shows demographic information for the full cohort at baseline, and the number of participants who returned for subsequent study visits. The most common reason for failing to participate in further visits was due to worsening of HD symptoms, with control participants most unlikely to return if their partner withdrew from the study or if they were experiencing relationship issues with their partner. During the course of TRACK-HD, 19 pre-HD participants converted to manifest HD as defined by having a new Diagnostic Confidence Score (DCS) score of ≥4, as described in section 1.1.3 (Huntington Study Group, 1996). At each visit, participants underwent a comprehensive battery of neuroimaging, cognitive, oculomotor, quantitative motor and neuropsychiatric testing. Many of the participants who were in TRACK-HD followed on to complete TrackOn-HD (see section 2.1.3).

2.1.3. TrackOn-HD

TrackOn-HD began in 2012 with a focus on the pre-HD phase of the disease process and included 3 visits, conducted at yearly intervals. At baseline, TrackOn-HD had 112 controls, 110 pre-HD and 21 early HD participants. Of these, 79 controls and 102 pre-HD participants were previous participants of TRACK-HD (with some pre-HD participants now classed as HD participants). An additional 33 controls and 30 pre-HD participants were also recruited.

Selection criteria were the same as for TRACK-HD, but participants who had previously taken part in TRACK-HD and who were older than 65 were able to enrol in TrackOn-HD. Demographic information for the baseline of TrackOn-HD is provided in Table 2.2. The range of assessments completed in TrackOn-HD was similar to that of TRACK-HD, but with more advanced imaging techniques, including task and resting state fMRI and diffusion imaging, performed at baseline, 12 and 24 months (See Figure 2.1). At 24 months two novel imaging sequences, Neurite Orientation and Dispersion Diffusion Imaging (NODDI; Zhang et al. 2012) and Chemical Exchange Saturation Transfer (CEST; Jones et al. 2006), were introduced. There were also some minor changes made to the cognitive, motor and neuropsychiatric batteries.

2.1.3.1. TrackOn-HD data collection

For the TrackOn-HD 2014 time point, data collection was performed as part of the work contributing to this thesis. Cognitive data collection and scoring for all participants who took part in the study at London was performed by myself as was over half of the MRI scanning (split with another team member).

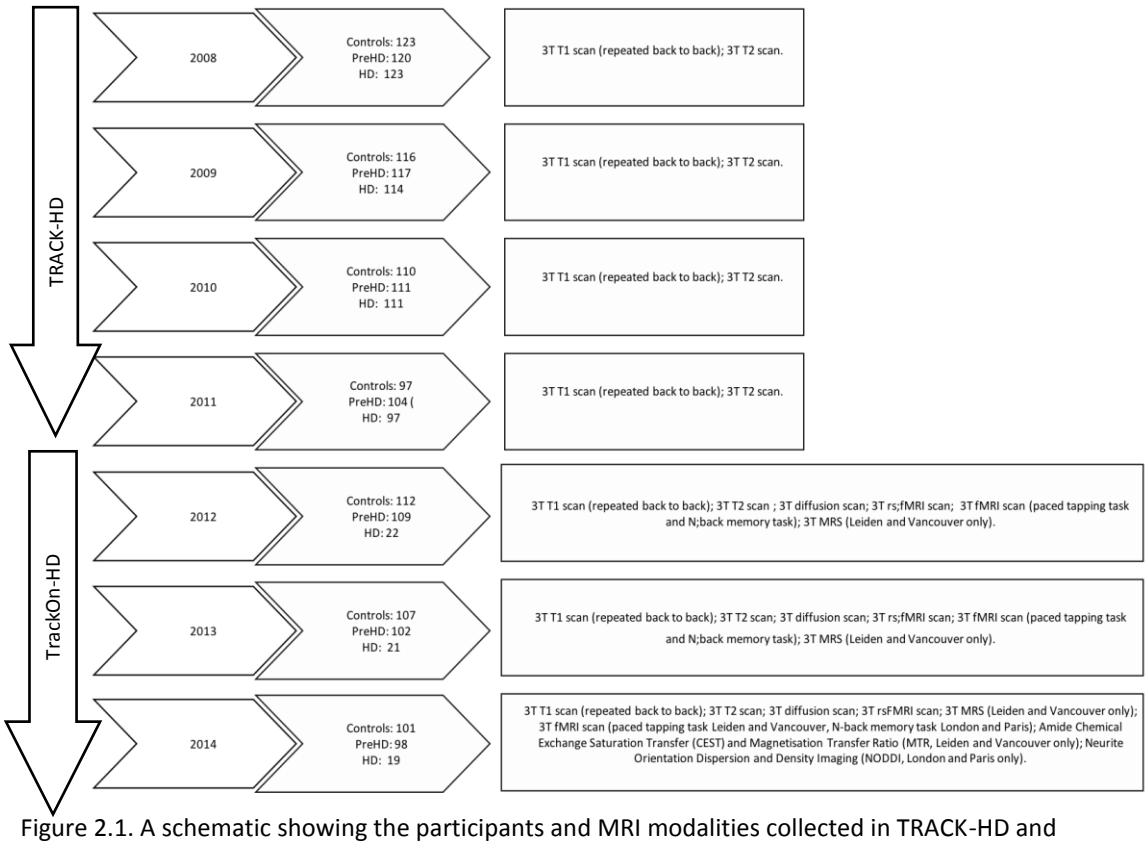


Figure 2.1. A schematic showing the participants and MRI modalities collected in TRACK-HD and TrackOn-HD.

Table 2.1 Demographic information for the TRACK-HD cohort at baseline. Data are mean (SD, range) or number (%). Disease-burden score=age x (CAG length - 35·5). Table adapted from Tabrizi et al. (2009).

	Controls (n=123)	preHD-A (n=62)	preHD-B (n=58)	HD1 (n=77)	HD2 (n=46)
Age (years)	46·1 (10·2, 23·0–65·7)	41·1 (8·6, 18·6–59·4)	40·6 (9·2, 22·3–64·1)	47·2 (10·3, 22·8–64·1)	51·4 (8·6, 33·3–63·3)
Women	68 (55%)	33 (53%)	33 (57%)	46 (60%)	21 (46%)
Education (ISCED score)	4·0 (1·3)	4·1 (1·1)	3·8 (1·3)	3·8 (1·3)	3·2 (1·4)
Disease-burden score	-	259·1 (30·1)	333·1 (30·0)	364·1 (74·3)	397·6 (67·5)
Site (n)					
Leiden	30	16	14	16	14
London	30	14	16	19	11
Paris	30	14	16	26	4
Vancouver	33	18	12	16	17
Time point (n)					
Baseline	123	62	58	77	46
12 Month	116	62	55	71	43
24 Month	110	60	51	70	41
36 Month	97	58	46	66	31

Table 2.2. Demographic information for the TrackOn-HD cohort at baseline. Data are mean (SD, range) or number (%). Disease burden = (CAG – 35.5) x current age.

	Controls (n=112)	preHD (n=110)	HD1 (n=21)
Age (years)	48.14 (10.70) 25.41-67.52	42.88 (9.10) 22.62-68.29	44.92 (7.96) 32.14-63.66
Women	67 (60%)	55 (50%)	14 (67%)
Education (ISCED score)	3.94 (1.04) 1-6	4.02 (1.00) 2-6	3.74 (1.01) 2-5
CAG	-	42.93 (2.27) 39-50	44 (2.59) 39-50
Disease-burden score	-	302.12 (53.04) 178.91-443.00	364.23 (56.69) 222.81-466.02
Site (n)			
Leiden	28	24	8
London	29	27	5
Paris	29	30	5
Vancouver	26	29	3
Time point (n)			
Baseline	112	110	21
12 Month	105	101	20
24 Month	101	99	18

2.1.4. PADDINGTON

PADDINGTON ('*Pharmacodynamic Approaches to Demonstration of Disease-Modification in Huntington's Disease by SEN0014196*') was a follow up from TRACK-HD with a focus on manifest HD patients. The data used in this thesis comes from work package two (WP2), which was a subcomponent of the PADDINGTON study designed to further validate imaging measures for use in short-interval clinical trials (Hobbs et al. 2013). There were 40 controls and 61 early HD participants in WP2. For this study there was a visit at baseline, 6 months and 15 months.

Participants were recruited across four sites: Leiden, London, Paris and Ulm. HD participants were required to have had a positive genetic test (CAG>36), be able to tolerate an MRI scan and sample donation and have no clinically significant and relevant history that could affect their involvement in the study. Control participants were mostly spouses or siblings of the HD participants who were tested negative for the expanded HD gene. Full selection criteria for the Paddington study is available (Hobbs et al. 2013). All participants gave written informed consent in accordance with the Decleration of Helsinki.

2.2. MRI Imaging

2.2.1. Acquisition

2.2.1.1. BrainWeb

The BrainWeb data is constructed from real MRI data, which was registered (aligned; see section 2.3.1), segmented and submitted to an MR simulator to create a 'phantom' version of the data. To create the 20 phantoms, 20 participants were scanned. All participants were scanned on a 1.5T Siemens Sonata Vision scanner. The T1 scan had 1mm isotropic voxels, with a 3-D spoiled gradient echo sequence (TR=22ms, TE=9.2ms). The scan was repeated four times to increase the signal-to-noise ratio and these four scans were averaged. All native images were corrected for intensity nonuniformity N3 bias correction (Sled et al. 1998). Registration was performed by randomly selecting one T1 from each subject and linearly registering this to the International Consortium for Brain Mapping (ICBM) average space. The remaining T1 images were linearly registered (rigid body, six parameters) to the first native T1, and from native space they were then registered to ICBM space. Finally, the volumes for each participant were averaged to create one image.

Next, each of the 20 average images was classed into 11 tissue classes; GM, WM, CSF, skull, marrow, dura, fat, fat2, muscles, skin/muscle and vessels (Aubert-Broche et al. 2006). The voxel intensity modelled tissue contributions within the voxel ranging from 0 to 1. The fuzzy volumes (PVE) defined the digital phantom. The next step was running the MR simulations to generate realistic 3-D images using both the PVE volumes from the anatomical phantom and parameters related to tissue, scan and coil aimed at making the simulation as realistic as possible. The MR simulator used a three step process. Briefly, the initial step was calculation of the signal intensities for different tissue types. Following this, the PVE volumes were used to weight the tissue intensities for each tissue type. The last step was to correct the images based on scan-dependent effects including noise and inhomogeneity.

Figure 2.1. shows the simulated MRI of one subject, alongside their real MRI and the phantom data on which the simulation was based. It should be acknowledged that there are discrepancies between the simulated data and the real data indicating that simulated data is not a perfect representation of a real MRI scan.

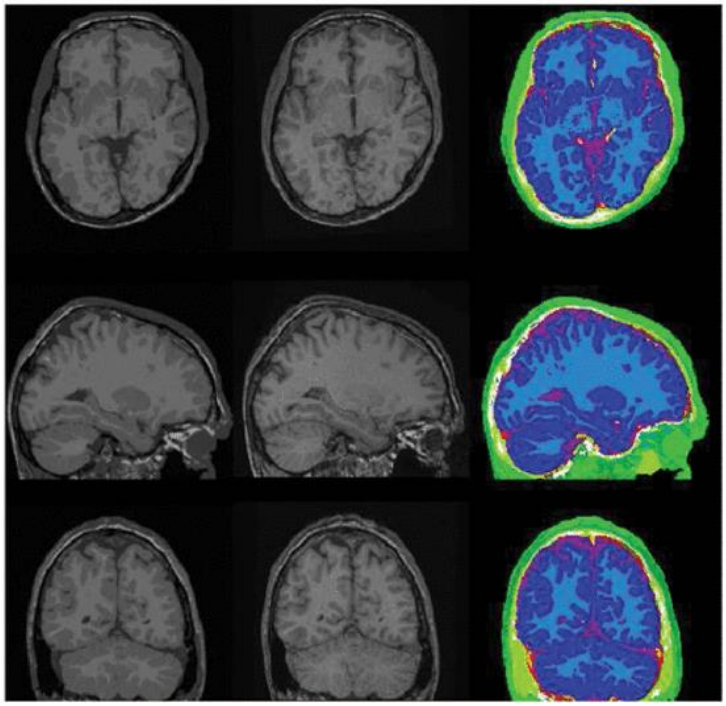


Figure 2.2. Simulated data (left), real data (middle) and phantom data (right) for one subject from the BrainWeb dataset (Aubert-Broche et al. 2006).

2.2.1.2. TRACK-HD, TrackOn-HD and PADDINGTON

3T T1-weighted scans were acquired from four scanners for both TRACK-HD and TrackOn-HD. Two were Siemens and two were Philips scanners. The parameters for Siemens

were TR = 2200ms, TE = 2.2ms FOV = 28cm, matrix size = 256x256, 208. For Philips TR = 7.7ms, TE = 3.5ms, FOV = 24cm, matrix size = 242x224, 164. The acquisition was sagittal to cover the whole-brain. There was a slice thickness of 1mm, with no gap between slices. These acquisition protocols were validated for multi-site use (Tabrizi et al. 2009). All images were visually assessed for quality; specifically artefacts such as motion, distortion and poor tissue contrast (IXICO Ltd. and TRACK-HD imaging team, London, UK). Acquisition protocols were the same for both studies.

For PADDINGTON, the scanners and imaging acquisition protocols were matched to those used in TRACK-HD for Leiden, London and Paris, for Ulm a Siemens Allegra scanner was used with slightly different acquisition paremeters, however this data will not be used in this thesis.

2.2.2. Pre-processing

2.2.2.1. BrainWeb

BrainWeb scans and associated GM regions were downloaded from http://brainweb.bic.mni.mcgill.ca/brainweb/anatomic_normal_20.html and converted from minc file format to Neuroimaging Informatics Technology Initiative (NIfTI) format using the visualisation software MIPAV (McAuliffe et al. 2001). The GT segmentations provided by BrainWeb were initially of a different resolution to the native space simulated T1 images. They were resampled via FSLmaths (Jenkinson & Smith 2001; Jenkinson et al. 2002). All outputs were visually checked to ensure successful resampling. Images were processed in NIfTI format.

2.2.2.2. TRACK-HD and TrackOn-HD

Images underwent initial quality control (QC) consisting of meta-data checks and visual QC to ensure that scan parameters were correct and to ensure no gross artefacts or errors were present, including motion artefacts and incorrect field of view placement (IXICO, London, UK). Examples of motion artefact are shown in Figure 2.3. Image data were archived at the Laboratory of Neuroimaging (LONI), University of California Los Angeles. Images from all sites were then downloaded from the LONI website for processing by the UCL TRACK-HD imaging team.

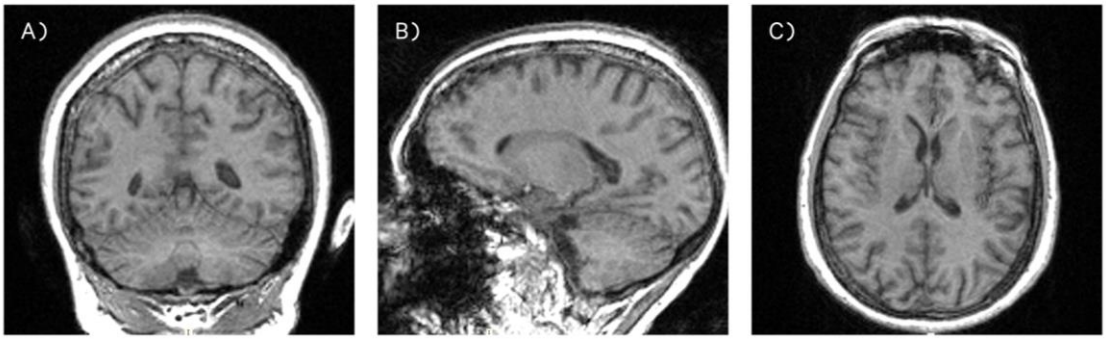


Figure 2.3. An example of a T1 scan from the TRACK-HD study that failed visual quality control because of a significant movement artefact.

Once downloaded, images were converted from dicom format to analyse format. MR images suffer from an artefact called intensity non-uniformity. This refers to the varying signal intensity that occurs across an MRI image, meaning that tissues of the same class have differing signal across T1-weighted scans. While this artefact has little impact on the ability to visually interpret an MRI scan, it can result in the failure of software to accurately delineate different tissue types. This is especially true for automated tools. Tissue non-uniformity is caused by a number of factors including radiofrequency coil homogeneity, local flip angle variations and participant anatomy (Sled et al. 1998; Boyes et al. 2008). To correct for this, images undergo a process called bias correction. This is a post-scanning process that calculates and corrects for the bias field within an image, Figure 2.4 shows a T1 scan both before and after bias correction. In the raw data, regions of high intensity can be seen in the medial sections of the scan. The bias corrected data appears to have more uniform tissue contrast across the brain. The non-parametric non-uniform intensity normalization (N3) method was used for all images in TRACK-HD, TrackOn-HD and PADDINGTON (Sled et al. 1998), with optimised parameters for 3T data as outlined in Boyes et al. (2008). Following conversion and bias correction, images were again examined by the team at UCL, and excluded from analysis if artefacts that were expected to affect analysis were present. Table 2.3 shows how many scans were excluded due to motion at each site for both TRACK-HD and TrackOn-HD.

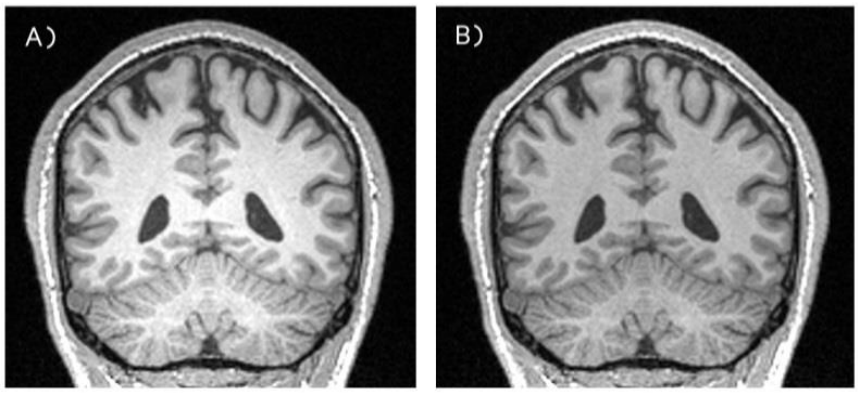


Figure 2.4. An example of a T1 scan (A) before and (B) after undergoing N3 bias correction, taken from the TrackOn-HD study.

Table 2.3. The number of scans at each time point that failed due to the presence of severe motor artefacts for the TRACK-HD and TrackOn-HD studies.

	2008	2009	2010	2011	2012	2013	2014
Number of scans failed due to motion	7	11	12	6	3	5	1
Total number of scans acquired	358	332	316	286	240	226	211

2.3. Structural MRI analysis tools

Measurement of different brain characteristics from sMRI can be performed using a number of different tools and techniques, including registration, segmentation and quantification of results. Throughout this thesis there are a number of methods consistently used to process T1 MRI scans. The most common methods are described here.

2.3.1. Image space and registration

MRI scans can be processed in a range of 'spaces', with a space referring to the position of the head within the 3D field of view (FOV). Native space is the space in which the scan was acquired, with the FOV and head in the position that the radiographer collected the data. Native space will be different for each scan collected, even for scans within the same scanning session if the FOV position was not copied between scans or the participant moved between scans. Standard space refers to a number of commonly used positions that are widely used across neuroimaging, which provide a common space to align different datasets to. The Montreal Neurological Institute space (MNI) is the most commonly used standard space. In addition, it is also possible to create a group template by concatenating all scans used in a particular study into a single study-specific template.

Typically, scans are aligned to a common space when performing cross-sectional group comparisons as this enables the direct comparison between different groups and participants. To align the position of a brain within the FOV to standard or template space, a registration needs to be performed. This is a process of alignment whereby a scan is moved from one space to another. Registrations can also be used to help quantify longitudinal change, especially for volumetric measures. During registration, Jacobian maps can be output and these can be used to quantify change. A Jacobian map is a 3D representation of voxel-wise contraction or expansion that occurs between each registered scan, and is based on the voxel-intensity of two scans. By multiplying the Jacobian map by the voxel-wise tissue maps output from a segmentation (section 2.3.2), change within each voxel can be quantified.

There are different types of registration and a number of different tools to perform these registrations, yet all registrations consist of two steps: a) calculation of registration parameters, b) application of these parameters to a scan.

2.3.1.1. Linear registrations

Linear registrations aim to achieve an alignment between scans, yet are restricted in the number of deformations that are performed. They aim to achieve a point-to-point mapping of two scans, however they will not correspond exactly especially when registering two brains from a longitudinal study that have undergone atrophy over time. Linear registrations perform the same action on all voxels and thus the content of the voxels themselves are not changed. During a linear registration, an MRI scan can be transformed via rotations, translations, shears and zooms, but not by the non-linear warping of the scan. The most basic type of linear registration is a rigid-registration, which uses 6 degrees of freedom allowing for three rotations and three translations to align the moving image to the target. Affine registration uses 12 degrees of freedom, which allows for three translations, three rotations, three shears and three zooms and enables more advanced co-registration of the scans. However, it may also result in some loss of information, particularly when registering atrophied brains that may show change over time.

2.3.1.2. Non-Linear registrations

Non-linear registrations allow for the scan to be warped in many dimensions, and they aim to achieve an exact correspondence between two or more images. A non-linear registration often initially starts with a linear registration step to roughly align the images, and then proceeds to a more detailed registration involving a large number of degrees of freedom

and warping in any direction. There are a number of types of non-linear registrations, with three methods used in this thesis described here.

2.3.1.2.1. Fluid registration

Fluid registration uses a model of a compressible viscous fluid to calculate the deformation (Crum et al. 2001; Freeborough & Fox 1998). Fluid models are typically performed on within-subject data, between two longitudinal scans consisting of a baseline and follow-up time points. This type of registration provides a good approximation of the gradual change seen in degenerating brains, whilst maintaining scan topology. Fluid registration is an iterative process that warps the follow-up scan to the baseline scan within the confines of the fluid model, while maximizing the correspondence within voxels of the two images. The result of a fluid registration is a follow-up scan that should correspond exactly to the baseline scan, along with a voxel-wise Jacobian deformation field mapping the transformations applied to each voxel. From this deformation field voxel compression maps can be calculated which enable visualization of voxel-wise expansion or contraction across the brain. In addition, volumetric change can be calculated between two scans by multiplying the baseline volumetric region by the Jacobian deformation field to quantify change. This is the method employed throughout TRACK-HD and TrackOn-HD studies to quantify regional change (Tabrizi et al. 2012; Tabrizi et al. 2013; Tabrizi et al. 2011).

2.3.1.2.2. DARTEL

Diffeomorphic Anatomical Registration Through Exponential Lie Algebra (DARTEL; Ashburner 2007) is a method of non-linear registration and template creation applied for between-subject normalization. It is included as part of the SPM toolbox, from version of SPM5, to improve between-subject registration prior to performing VBM (section 2.3.4.2). By using DARTEL, the scans from different participants are not only aligned but also used to create a study-specific template. When registering pathological brains to a standard template, errors can be introduced since the brains do not achieve good correspondence. By creating a study-specific template these biases and errors can be reduced. DARTEL aims to create a continuous and smooth one-to-one mapping that is invertible (reversible) and maintains topology.

Images are first segmented into GM, WM and CSF tissue classes using SPM segment (section 2.3.4.2). During this process, the images are roughly aligned and deformation parameters are saved. When performing DARTEL, the tissue classes are first combined using the deformation parameters to create a mean image used as an initial template. Deformations

between this rough template and the individual images are computed, and the template is re-generated using the inverse deformations to the images, and this is averaged. This is repeated 7 times (by default) to iteratively improve the template. The output is a GM and WM template, along with flow fields that contain the parameters to warp between the native scans and the template.

2.3.1.2.3. Longitudinal serial registration

A further registration that is applied in the current thesis is longitudinal serial registration. When performing quantification of atrophy across time points, it is important to ensure that between-scan noise and bias are reduced as much as possible. Previous methods of longitudinal registration were criticized for introducing additional bias by the process of registering all follow-up scans to the baseline scan (Ashburner & Ridgway 2012). It was suggested that the process of applying a non-symmetrical registration pipeline could result in the detection of false positive differences. Similarly to DARTEL, this technique uses generative modelling to create an average of multiple scans. However, here the scans are within-subject. Along with registering the scans and creating an average, the approach also includes an integrated correction for between-scan (differential) intensity homogeneity (Ashburner & Ridgway 2012). The method first applies a rigid-body transformation to ensure scans are in rough alignment, and then enters into a process of between-scan diffeomorphic iterative registration to create an average image and deformation parameters. These parameters can then be used to calculate longitudinal volumetric change across multiple time points. The method has been validated previously, and was shown to be a consistent method with biologically plausible results (Ashburner & Ridgway 2012).

2.3.2. Partial volume effects

The partial volume effect (PVE) refers to the inability of MRI to represent the true proportion of each tissue type within a voxel due to the resolution of voxels (Tohka 2014). For example, a 1x1x1mm voxel may contain only GM or a proportion of WM or GM. Partial volume of voxels is typically accounted for by giving each voxel a likeliness of belonging to each tissue type ranging from 0-1. Most segmentation techniques calculate probabilistic segmentation voxel-wise maps, which account for partial volume effects, whereby different tissue maps are created (e.g. for GM, WM, CSF) and for each tissue map every voxel is given a likelihood of belonging to that tissue ranging from 0-1. For example, we would expect a voxel within the middle of the lateral ventricles to be given a value of 1 for the CSF tissue map, but 0 for the

WM tissue map. Many tools also output 'discrete' volumes whereby each voxel is assigned to a particular tissue type in a binary 1 or 0 method. It is widely accepted that partial volume maps result in more accurate quantification of tissue volumes (Tohka 2014). Figure 2.5 shows examples of partial volume maps and discrete tissue maps for the same participant.

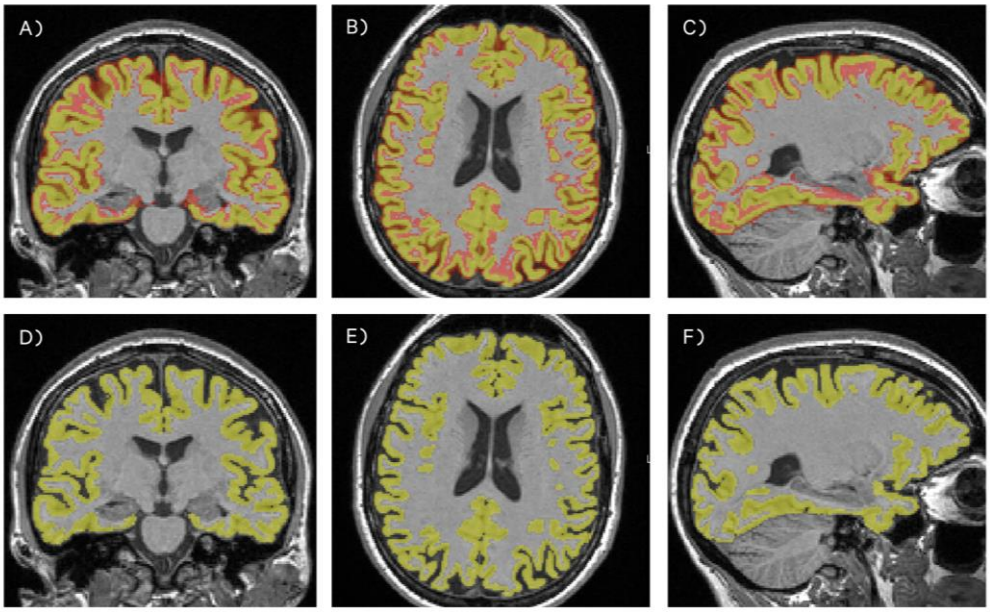


Figure 2.5. A scan from the TRACK-HD study, showing a partial volume map of the cortex in (A) coronal, (B) axial and (C) sagittal views, and a discrete tissue map in (D) coronal (E) axial and (F) sagittal views.

2.3.3. Manual delineation of regions using MIDAS

Often considered a 'gold standard', manual delineation of brain regions can be performed to outline regions of interest and calculate the volume of these regions. This can be done in a variety of ways, but the method used in this thesis makes use of an in-house UCL software that has been used in a large number of studies examining brain volume in Alzheimer's Disease and HD. The Medical Image Display and Analysis Software (MIDAS; Freeborough et al. 1997) is an interactive software that enables 3D processing of MRI scans. It has a wide variety of options for viewing and segmenting scans. Most processes begin with an intensity-based thresholding step defined by a pre-specified mean-brain intensity, followed by manual editing of the region. The use of pre-defined mean-brain intensity values ensures that consistent delineation can be performed on a particular scan. The first step is typically to measure the whole-brain volume, and this can then be used to calculate thresholds for other regions. After whole-brain segmentation is performed, the brain region is registered to

MNI305 standard space. By registering the whole-brain region and scan to standard space landmark-based cut-off points used when segmenting the other regions can be applied consistently within subsequent scans for the same participant and between all participants.

The process of performing manual segmentation generally results in a high degree of reliability both within and between raters, and the regions tend to be visually accurate. However, it is also labour intensive and involves a rigorous period of training. In order to be qualified to segment images for a study within the HD Centre, a rater needs to pass a 'test set' of five scans, performing segmentation of the test set twice with a week in between the two repeated segmentations. The volumes for these five scans are required to be in line with set reliability thresholds, which are set at 1% difference for global measures, such as brain, total intracranial volume (TIV) and ventricle volumes, and 3% difference for regional segmentations such as the caudate. Despite the intensive training and high reproducibility, it is important to acknowledge that while raters are blinded to disease stage of the participants, it is sometimes possible to determine disease status from scans in particularly atrophied brains, introducing possible rater bias.

2.3.3.1. Whole-brain segmentation

Whole-brain segmentation is performed in native space, with the first step being the selection of an appropriate threshold that excludes most non-brain tissue via an interactive threshold-selection process. The rater then manually edits this region where required, with the aim to perform as little manual editing as possible in order to limit rater bias. Whole-brain segmentations take between 45-60 minutes. Manual delineation of the whole-brain is performed at baseline, with baseline regions then propagated (i.e. registered and applied) to follow-up time points and manually edited where necessary if the propagation has leaked into non-brain regions. Figure 2.6 shows an example of whole-brain segmentation.

To measure longitudinal atrophy, the Brain Boundary Shift Integral (BBSI) is calculated. This involves fluid registration of the follow-up scan and region to baseline, differential bias correction to adjust for any differences in intensity between the two scans, dilation of the baseline region, and quantification of the change in the boundary of the baseline region compared to the follow-up region (Freeborough & Fox 1997). Whole-brain segmentation has been demonstrated as a sensitive measure of group differences between pre-HD, HD and controls, and the BBSI has been found to be sensitive to longitudinal brain-atrophy in HD (Tabrizi et al. 2009; Tabrizi et al. 2012). It is currently being used in a number of on-going clinical trials for HD.

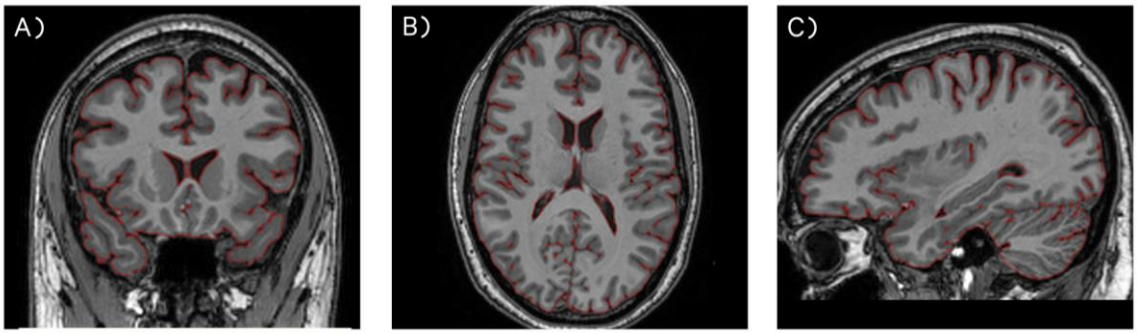


Figure 2.6. An example of a whole-brain segmentation from the TrackOn-HD study, shown in (A) coronal, (B) axial and (C) sagittal views.

2.3.3.2. Ventricle segmentation

Following whole-brain segmentation, the region is registered to standard space and ventricle segmentation is performed in standard space. A pre-specified upper intensity threshold of 60% of the mean brain intensity is used (measured from the previously segmented brain region). The lower threshold is set at the minimum possible value. Ventricle segmentation includes the lateral ventricles and temporal horn of the lateral ventricles, but not the third or fourth ventricles (Scahill et al. 2003). Figure 2.7 shows an example of a ventricle segmentation, which takes between 10-15 minutes to complete. The ventricles were segmented at each time point in TRACK-HD. The Ventricle BSI (VBSI) is used to measure ventricle expansion, with a similar method to the BBSI applied. Again, ventricular volume and ventricular expansion are sensitive to between-groups differences and longitudinal change in HD (Tabrizi et al. 2009; Tabrizi et al. 2012).

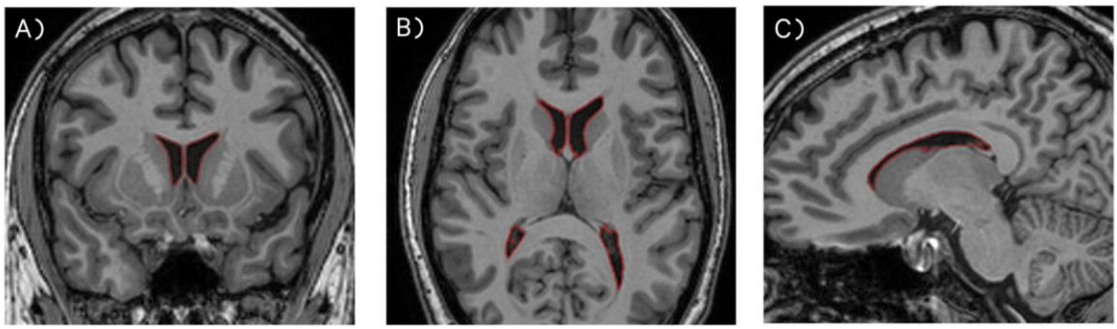


Figure 2.7. An example of a ventricle segmentation from the TrackOn-HD study, shown in (A) coronal, (B) axial and (C) sagittal views.

2.3.3.3. Caudate segmentation

Caudate segmentation was performed on the standard-space registered scan, and includes the head and body of the caudate. The intensity thresholds are set at lower and upper values of 62% and 111% of the mean brain intensity respectively (Hobbs et al. 2009). The caudate regions generally require more manual editing than whole-brain and ventricular regions due to the reduced contrast between the caudate and WM. An example of a caudate segmentation is shown in Figure 2.8, and segmentations take between 60-90 minutes. The caudate was only segmented at baseline, and the Caudate BSI (CBSI) was used to measure caudate atrophy (Hobbs et al. 2009). Caudate atrophy is one of the most significant measures of structural volume change in HD (Tabrizi et al. 2013).

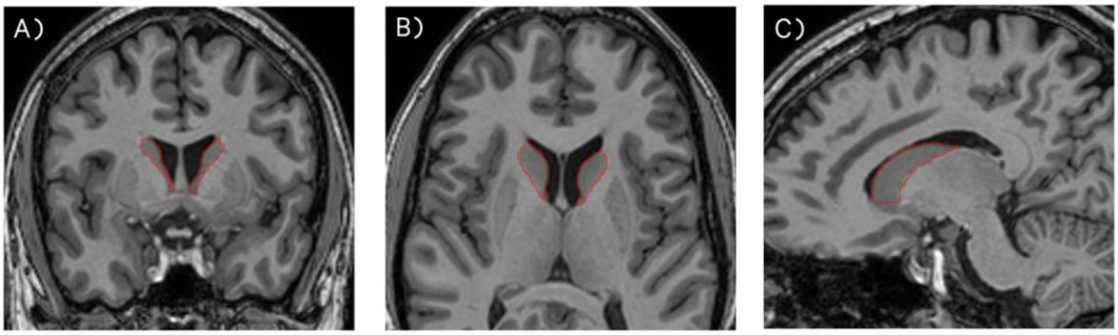


Figure 2.8. An example of a caudate segmentation from the TrackOn-HD study, shown in (A) coronal, (B) axial and (C) sagittal views.

2.3.3.4. Putamen segmentation

Similarly to caudate segmentation, segmentation of the putamen is performed on the whole-brain registered to standard space. While this protocol has not been published, it underwent similar validation to the other methods cited here and has been applied in a number of research studies and HD trials (Hobbs et al. 2013; Hobbs et al. 2015). Intensity thresholds are set to 90% and 112% of the mean brain intensity, and the region extends superiorly until the last slice clearly containing putamen and inferiorly until the last slice in which the putamen is clearly separated from the caudate by the internal capsule. An example of a putamen segmentation is shown in Figure 2.9, and segmentations take between 60-90 minutes.

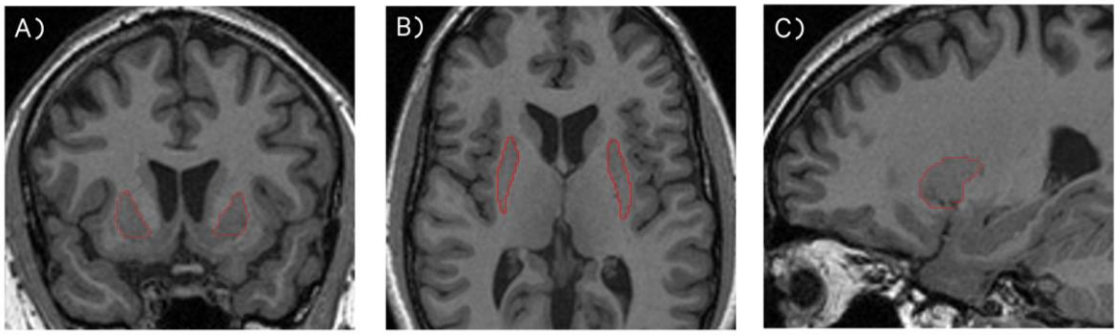


Figure 2.9. An example of a putamen segmentation from the PADDINGTON study, shown in (A) coronal, (B) axial and (C) sagittal views.

2.3.3.5. Total intracranial volume segmentation

Total intracranial volume (TIV) is a measure of the total volume within the skull, including brain, meninges and CSF. TIV is regularly used to control for inter-subject variability in head size (see section 2.4.3), since it is associated with total and regional brain volumes (Malone et al. 2015). TIV is measured via first setting pre-specified thresholds of a lower threshold of 30% of the mean brain intensity, and an upper threshold at the maximum possible value. Starting at the base of the cerebellum, one slice every 10 slices is seeded resulting in selection of the outside of the dura, as shown in Figure 2.10. These regions are then filled and propagated, to create a region that fills the whole intracranial space. TIV segmentations take between 10-15 minutes. This measure has been demonstrated as highly reliable and preferable to performing automated quantification of TIV (Malone et al. 2015).

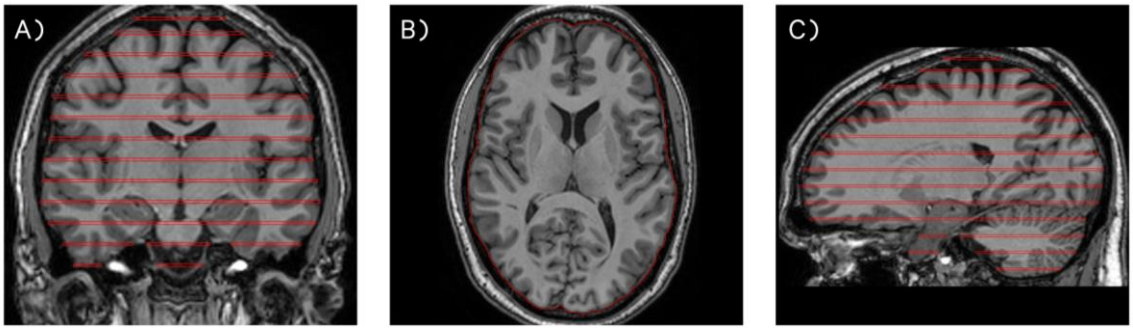


Figure 2.10. An example of a TIV segmentation from the TrackOn-HD study, shown in (A) coronal, (B) axial and (C) sagittal views. The slices are filled to calculate volume within the region.

2.3.3.6. Role within TrackOn-HD data processing

For the TrackOn-HD 2013 visit I performed T1 data processing. This included data download, conversion and bias correction, visual QC and manual segmentation of the whole-

brain and ventricles for all 2013 scans (see section 2.3.3). Furthermore, I performed retrospective segmentations for the 2012 time point, including all manual TIV and caudate segmentations, along with some brain and ventricle segmentations. Segmentations for other time-points and regions were performed by the TRACK-HD, TrackOn-HD and PADDINGTON imaging teams.

2.3.4. Automated Delineation of Regions

In addition to manual segmentation, there are a number of techniques that automatically delineate different brain regions. These techniques have varying requirements and can perform a number of different measurements, including measures of volume, CT and surface-based characteristics of the brain such as curvature. The measurement of volume and CT using these techniques is covered in more detail in Chapters 3 and 4, however a brief description of the most common methods referred to in this thesis are below.

2.3.4.1. FMRIB Software Library

The FMRIB Software Library (FSL; Jenkinson et al. 2012; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) is a suite of tools for functional, structural and diffusion MRI analysis and statistical analysis. Within FSL, the FMRIB's Automated Segmentation Tool (FAST) is frequently used to segment different regions of the brain. FAST corrects for inhomogeneity within the images and uses a hidden Markov random field model and an associated expectation-maximisation algorithm to segment an image into tissue types (Zhang et al. 2001). The segmentation process generates tissue classes (including GM, WM, CSF), and once the tissue classes are segmented masks can be applied to calculate volumes from sub-regions of these tissues.

2.3.4.2. Statistical Parametric Mapping

The Statistical Parametric Mapping (SPM; http://www.fil.ion.ucl.ac.uk/spm/) programme is another tool that offers analysis options for structural, functional and diffusion data, as well as MEG and EEG data. It is frequently used in HD studies to perform volumetric segmentation, and was used throughout the TRACK-HD and TrackOn-HD studies to measure WM and GM total volume (Tabrizi et al. 2011; Tabrizi et al. 2009; Tabrizi et al. 2012; Tabrizi et al. 2013). Once segmentation has been run on an image, the tissues can also be used to perform a VBM analysis, see section 2.3.4.2. Tissue segmentation involves modelling of intensity distributions within the scan via Gaussians and tissue probability maps, with priors

used to weight classification (Ashburner & Friston 2005). This approach combines normalisation and segmentation. The two most commonly used versions of SPM currently in use are SPM8 and SPM12. SPM8 has two segmentation procedures, Unified Segment and New Segment. SPM12 has one main procedure, Segment, although it also provides access to Unified Segment.

Unified Segment was first released in 2005 (Ashburner & Friston 2005) and it combines tissue classification, bias correction and image registration to segment the GM, WM and CSF via a voxel-wise approach. The tissue classification phase uses intensities to classify voxels into a tissue category. The intensity distributions are modelled by a mixture of Gaussians, and tissue probability maps (TPMs) registered to standard space (ICBM152) are used to improve classification (Ashburner & Friston 2005). The TPMs give the prior probability of any voxel being a member of GM, WM, or CSF. To align the TPMs with each individual brain, initial affine registrations are performed followed by non-linear registration to improve the fit, and the inverse of these registrations is used to align each brain to standard space. The output tissue values range from 0-1 and represent the probability that each voxel belongs to a certain tissue class.

The New Segment toolbox is an extension of Unified Segment. This method is designed to treat the mixing proportions differently, use an improved registration model and process multi-spectral data (i.e. both T1 and T2 images). It also has an extended set of tissue probability maps allowing for voxels outside of the brain to be treated differently. The segmentation option in SPM12 is a modified version of New Segment, with some new default options.

SPM is freely available, although it does require Matlab (which is not freely available). It is widely used within the neuroimaging field.

2.3.4.2.1. Voxel Based Morphometry

VBM is a voxel-wise approach to examining brain differences in sMRI. By registering the MRI scans for all participants to a template and then smoothing the scans, differences in brain volume can be compared between groups or across time within individuals at each voxel. VBM allows for examination of volumetric differences without specification of ROIs and can detect small regions of difference that may be overlooked when using larger regions to compare between groups. Different software can be used to perform VBM analyses, but here the application of SPM VBM is covered (Ashburner & Friston 2000). To perform VBM within SPM, images are first segmented in native space using SPM Segment as described previously. A group template is also created using the DARTEL toolbox (Ashburner 2007). The segmented

regions are then warped to the DARTEL template and modulated and smoothed to account for any changes that occurred during normalisation to the template. Following this, a statistical model is specified and group differences can be compared across the whole-brain. Regions of interest can also be specified using a mask, if there are *a priori* hypotheses about the location of group differences.

2.3.4.3. BRAINS

BRAINS (Brain Research: Analysis of Images, Networks, and Systems; Magnotta et al. 2002) is another freely available automated technique that utilizes data from both T1- and T2-weighted scans to perform brain segmentation. It uses a neural net trained on human rater definition of the putamen and atlas-based structure identification, followed by an additional boundary correction to ensure no structural overlap. While BRAINS has been used in a number of studies investigating the brain in HD (Nopoulos et al. 2011; Paulsen et al. 2006), including the TRACK-HD study (Tabrizi et al. 2009; Tabrizi et al. 2011; Tabrizi et al. 2012; Tabrizi et al. 2013), it is not widely used.

2.3.4.4. FreeSurfer

FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) is a widely used software that aims to segment, measure and analyse neuroimaging data. It has a volume-based stream and a surface-based stream. The volume-based stream aims to process MRI volumes and classify subcortical regions (Fischl et al. 2002; Fischl et al. 2004). The brain is affine registered to Talairach space (a frequently used standard space), and then an initial volumetric labelling step is performed followed by intensity correction. Following this, a high dimensional nonlinear volumetric alignment to a Talairach atlas is performed. In the final phase the volume is labelled. The volume-based stream has evolved separately to the surface-based stream and a number of the stages are different (i.e. the registration and bias field correction). The surface-based stream measures features such as CT, surface area and curvature of the brain. Briefly, after correction for field inhomogeneity, affine-registration and skull stripping, the WM voxels are classified based on their intensity, their neighbours' intensity, and their position. A triangular mesh is built around the WM, which is then smoothed and topologically corrected. The external cortical surface is created via expansion of the WM surface until it reaches a point of maximal contrast between the GM and CSF.

FreeSurfer is widely used, and has been applied within HD research a number of times, although mostly for CT analysis or to use regional volumes as a seed region or an extraneous

variable in analyses (Rosas et al. 2008; McColgan et al. 2015). There are standard pipelines available and FreeSurfer is easy to implement.

2.3.4.5. Advanced Normalization Tools (ANTs)

ANTs is a more recently developed tool, and while the segmentation method has not been used in neurodegeneration research to date it is freely available and gaining popularity within the neuroimaging community. A fully automated stream that outputs volumetric measures and CT is available (Diffeomorphic Registration-based Cortical Thickness, DiReCT) along with a separate volumetric stream (Atropos; Das et al. 2009; Tustison et al. 2014). Firstly N4 bias correction (Tustison et al. 2010) is performed on the raw MRI image, followed by brain extraction. The brain segmentation uses a mixed segmentation and template-based extraction. Following this, tissue segmentation is performed via alternating between prior-based segmentation and N4 bias correction. To calculate CT a prior-constrained estimate of the distance between the GM/WM boundary and the GM/CSF boundary is used. ANTs pipelines are freely available and online support is provided.

2.3.4.6. Multi-Atlas Label Propagation with Expectation-Maximisation based refinement (MALP-EM)

MALP-EM is another recently developed tool aimed at providing a fully automated segmentation method able to cope with severe brain pathologies (Ledig et al. 2015). It utilises a previously described registration approach (Heckemann et al. 2012), atlas-based segmentation, and intensity-based expectation maximisation to segment the brain into cortical and subcortical regions (Ledig et al. 2015). MALP-EM has been validated on traumatic brain injury patients and is able to stratify patients into favourable vs. unfavourable outcomes based on MALP-EM segmentations (Ledig et al. 2015). It is freely available and has recommended usage included.

2.4. Statistical Analysis

Different statistical analyses were used throughout this thesis. This section covers some of these methods, with further information provided where relevant later in the thesis. All statistical analyses were performed within STATA version 12.1, or within a neuroimaging analysis tool such as SPM.

2.4.1. Measures of accuracy and reliability

Chapter 3 of this thesis aims to compare various tools that perform automated segmentation of MRI scans. Measuring the reliability of these tools is a recognised challenge (Bouix et al. 2007). If two tools produce volumetric values for a tissue type that correspond, there is no guarantee that they spatially overlap. Because of this, both real study MRI data from TRACK-HD and phantom data with corresponding GT segmentations were used in this thesis to assess the accuracy and reliability of segmentation tools. These statistical methods are described here, with implementation discussed in Chapter 3.

When using data such as the BrainWeb dataset, which has a corresponding GT segmentation for each scan, overlap scores (the dice coefficient) can be calculated on a voxel-wise basis as discussed in Crum, Camara and Hill (2006). The dice coefficient is a measure of spatial overlap and has been used to validate segmentation techniques in a number of studies previously (Zou et al. 2004). The technique developed by Crum et al. compares the voxel-map output by each segmentation tool to the corresponding GT in a voxel-by-voxel method to determine the overlap of whole images. It accounts for partial volume estimates thus enabling a precise measure of spatial overlap in MRI. In addition to spatial overlap, when using a GT measure of brain volume, paired t-tests can be performed to determine whether there are significant differences between the GT and a segmentation method. Finally, Bland Altman plots and Pitman's Test of difference measure the agreement between the GT and the output from a segmentation tool.

When calculating the reliability from back-to-back measurements, as also done in Chapter 3 on the TRACK-HD data, reliability can be measured using the intraclass correlation (ICC; Bartko, 1966). ICC measures the agreement between repeated ratings, and ranges from 0-1 with higher values demonstrating better reliability. Mean repeatability for back-to-back scans was also calculated as percent variability error (Function 1; Jovicich et al. 2013; Tustison et al. 2014). This measure provides a percentage value of variability between repeated applications on the same cohort for each tool; lower values represent less variation. Finally, Spearman's Rho was also used to test the correlation between each set of volumes extracted using each segmentation tool.

$$\varepsilon = \frac{(Vol_A - Vol_B)}{0.5 \times (Vol_A + Vol_B)} \times 10$$

Function 1: Mean repeatability calculation used on back-to-back scans from 2008. Vol_A represents the first baseline back-to-back scan, Vol_B Represents the second baseline back-to-back scan.

2.4.2. Multiple regression

Multiple regression uses a general linear model (GLM) to estimate the relationship between predictor variables (e.g. Huntington's disease stage) and measured (criterion) variables (e.g. GM volume). Multiple regression enables the specification of multiple criterion variables so the unique contribution of variables to a predictor can be understood. Multiple regression is used throughout this thesis to understand different relationships between variables.

2.4.3. Covariates

Covariates are confounding variables that may be related to a variable of interest, but are not of interest in themselves. They can be statistically controlled for during analysis, resulting in a more direct measurement of the relationship between variables of interest. All analyses in this thesis controlled for age and sex, due to the known effects of these two variables on brain size (Barnes et al. 2010) In addition, site was used as a covariate to control for the multi-site nature of TRACK-HD and TrackOn-HD with the aim of reducing variance due to site. However, it should be noted that this does not fully account for effects of site, as sites are differentially affected by artefacts such as scanner drift, software upgrades and field inhomogeneity. All participants were scanned at the same time of day. Education was used as a covariate when examining cognitive variables, due to the relationship between education level and cognitive performance (Beeri et al. 2006), however was not used as a covariate in other analyses. This was done to maintain consistency with previous TRACK-HD analyses, which do not control for education unless performing analyses that include cognitive variables. This decision was made by the TRACK-HD steering committee in early TRACK-HD meetings. It was widely discussed by the imaging and cognitive teams prior to a decision being made. Medication was not included as a covariate. Due to the large number of medications being taken by participants throughout data collection of the TRACK-HD and TrackOn-HD studies, along with the variable doses taken across time-points, the data could not be accurately controlled for and was thus excluded from the covariates.

TIV is a measure that has been frequently discussed as a potential covariate in neuroimaging literature (Malone et al. 2015; Barnes et al. 2010). TIV acts as a proxy to 'healthy' brain volume, since TIV measures the total intracranial space and has been demonstrated to remain constant despite increasing neural atrophy (Whitwell et al. 2001; Matsumae et al. 1996). By controlling for TIV, variables of interest can be compared whilst taking into account differences due to head size. The TIV can be measured via manual

delineation of the region (Section 2.3.3.5), which has been demonstrated a valid method of calculating TIV (Malone et al. 2015). This technique was used for all data in this thesis. TIV was controlled for by either using it as a regressor in the statistical model, or a by expressing volume as a percentage of TIV.

2.4.4. Correction for multiple comparisons

The results of statistical analyses are prone to both Type I and Type II errors. A Type I error refers to the incorrect rejection of a true null hypothesis (a false positive). A Type II error is the acceptance of the null hypothesis when a true effect is present (a false negative). The more statistical comparisons performed in a given analysis, the more likely a Type I or Type II error is to occur. While an understanding of these two scenarios is relevant to all researchers undertaking statistical analysis, the nature of neuroimaging analysis and the sheer volume of statistical comparisons conducted during many types of neuroimaging analyses mean that they are more likely to occur than within other fields (Lindquist & Mejia 2015; Hupé 2015). For example, when performing a group comparison using VBM via a GLM analysis, all of the voxels within a given region are statistically compared between groups to find regions of significant difference. This involves performing a univariate test at every voxel, resulting in thousands of comparisons.

By performing correction for multiple comparisons, a researcher reduces the likelihood of Type I or Type II errors occurring. Correction for multiple comparisons is typically performed by adjusting a statistical threshold based on the number of comparisons being performed, with the Bonferroni correction the most widely known (Lindquist & Mejia 2015). Within neuroimaging there are two commonly used methods of correction: correction for family-wise error rate (FWE) and correction for false discovery rate (FDR).

FWE refers to the likelihood of making one or more Type I errors in a family of tests (i.e. multiple tests), it includes the Bonferroni method. Correcting for FWE assumes that we want to limit the chance of a Type I error in any of our statistical tests, and is often too stringent, thus increasing Type II errors. FDR, in comparison, controls for the number of Type I errors across all significant results, and aims to reduce the number of false positives only within the subset of voxels found to be significant.

2.5. Dynamic causal modelling

It is known that while brain regions have functional specificity, they also interact with each other both functionally and structurally. The connections and strength of interaction between regions is dependent on the neural activity being performed. However, traditional

GLM methods and connectivity techniques used to analyse both sMRI and fMRI data cannot be used to estimate causal interactions between variable brain networks. Instead they are used to correlated volume or activity in different regions with each other or with external variables, such as task performance. Post hoc interpretations can then be drawn on the existence of inter-regional networks. Because of this limitation, a number of techniques have been developed to estimate models that enable the testing of hypothesised inter-connecting brain networks. In 2003 Dynamic causal modelling (DCM) was developed for this purpose (Friston et al. 2003). DCM is an analysis technique, typically used with fMRI or electrophysiological data, which aims to infer causal relationship between distributed neuronal populations. Different network models are hypothesised and tested, and Bayesian model selection (BMS) is then used to compare the different generative models in order to find the most appropriate model to fit the measured fMRI or electrophysiological data. Based on the strongest model, inferences can then be drawn on the biological structure of neuronal networks. Recently, the models and assumptions applied in DCM have been adapted for application to sMRI data in order to perform more powerful characterisations of both within-region and inter-regional atrophy than approaches such as VBM (Ziegler et al. 2017).

DCM was developed to quantify effective causality, that is, to examine the causal influence that neuronal systems exert over each other. The model estimates the relationships, referred to as coupling, between brain regions and enables examination of how the coupling is influenced by changes in the experimental context. DCM creates realistic models of the temporal evolution of a set of neuronal states (x). The change within a region is dependent on the current state of the region (x), external inputs (C) that cause network-wide changes and the context-dependent connections to other regions (A), and modulatory parameters (B) which are factors that modulate activity (Friston et al. 2003). These expressions are the main components of the basic DCM model, called the bilinear model, shown in Figure 2.11.

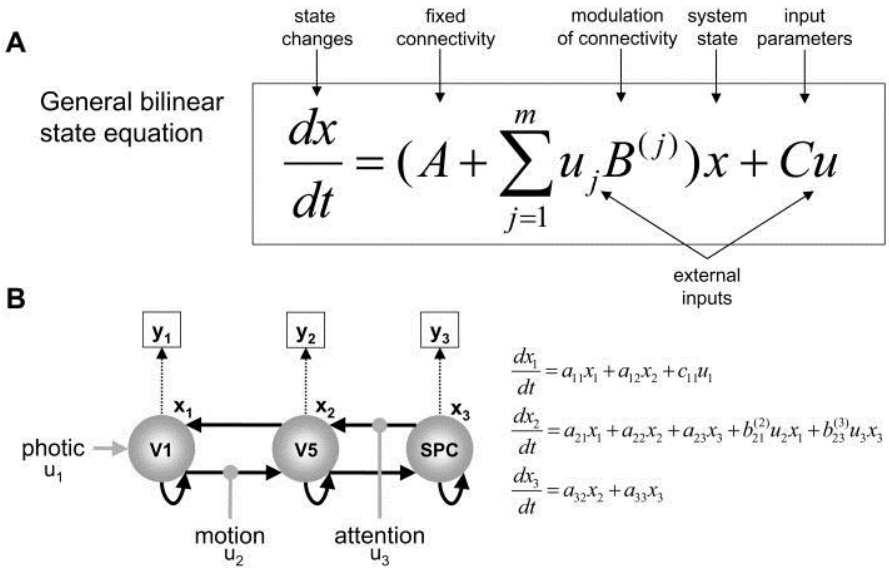


Figure 2.11. Shows (A) the bilinear state equation of DCM for fMRI, with an example (B) of the dynamics in a hierarchical visual system consisting of areas V1 and V5 and the superior parietal cortex (SPC). Each region has a state variable (x_1, x_2, x_3), and connections between regions are represented by the black arrows. The external inputs are represented by grey arrows, and dotted arrows indicate the transformation from neural states into haemodynamic observations. Visual stimuli (photic) drives activity in V1 which is then propagated to V5 and the SPC, with the V1-V5 connection changing when stimuli are moving, and the SPC-V5 connection is modulated by attention. The state equation for this scenario is shown on the right. Adapted from Stephan et al. (2007).

DCM combines the bilinear model with a validated hemodynamic model, which describes the process of a BOLD signal being generated from neuronal activity. The hemodynamic model defines the change in blood volume and deoxyhaemoglobin, the basis of an fMRI signal, that is predicted when changes in regional neural activity result in vasodilation and a subsequent increase in blood flow.

A model is specified based on the hypothesised hemodynamic model, and parameters are estimated such that both the modelled and measured BOLD signals reach maximal agreement. The model is grounded in priors that contain previous knowledge about the model, including the hypothesised hemodynamic response within regions but also the neural interactions between regions. The process of model estimation is iterative; the estimation is modelled based on the current parameters, results are then compared to the measured data, improvements are made on the model parameters and the cycle repeats. Different hypothesised models can be specified and estimated, with the final estimated models statistically compared. Bayesian model selection (BMS) is then used to determine the hypothesis that best explains the measured BOLD signal.

While this type of classical DCM is typically used in functional or electrophysiological data, recent work has adapted the framework to permit the application of similar methodology to sMRI data. The approach enables linear and non-linear modelling of longitudinal change in multiple structural regions, offering a unique approach to quantify structural change (Ziegler et al. 2017). Mass univariate methods of quantifying neural change typically involve whole-brain analyses, which can identify linear change associated with different correlates. The model by Ziegler et al. (Ziegler et al. 2017) uses similar principles as in DCM and the bilinear model to study structural changes in a systems-based perspective. This approach avoids issues of multiple comparison corrections that arise from mass-univariate methods, but also enables the comparison of multiple hypothesised models of neuronal change via BMS. Briefly, the model uses a hierarchical system to perform first- (individual) level and second- (group) level modelling. Similarly to DCM for functional data, a bilinear model is used that specifies the temporal progression of states, in this case volume, over time (multiple scanning sessions). The states are influenced by within-region connections (atrophy) and between-region influences, that is, the effect that regions have upon each other. The examination of between-region connections should be hypothesis driven since the model is specified by the user, and thus a good empirical basis for this model is essential. The trajectories of the volumetric states are also influenced by external factors, which can include known influences, such as genetic factors that may be related to neural atrophy and can be included in the model. External factors can also include unknown influences, for example unknown hormonal influences that could impact on atrophy.

First-level individual models can be applied and compared, for example a quadratic fit could be compared to a linear fit to determine which model might best explain the data. Following selection of a first-level model the model is applied to all data and then second-level models can be used to construct group-wise progression trajectories. The second-level model also incorporates covariates to account for individual differences in age, sex and other variables that are not of interest.

This analysis can be used to study within-region trajectories of change and also the interactions between regions, and can tolerate variability within time points and some missing data. This thesis uses the approach to undertake structural dynamical modelling of atrophy in the cortex in HD. See Chapter 6 for detailed information on the implementation of the model in this thesis.

3. COMPARISON OF AUTOMATED GREY MATTER SEGMENTATION TOOLS

Grey matter volume is a frequently used measure of brain morphology; it is typically reliable and able to discriminate between healthy controls and clinical groups (Schwarz et al. 2016). It is most often measured using automated software tools that separate GM from other tissue types. Thus, high quality delineation (segmentation) of GM, WM and cerebrospinal fluid (CSF) is critical in achieving accuracy for volumetric analyses. Recent work has demonstrated that variability between volumetric tools can result in inconsistencies within the literature, driving false conclusions about neurological conditions (Katuwal et al. 2016). There are a number of automated tools that can be used for performing GM segmentation, with each tool having differing characteristics. Methodological comparisons of these tools have focused on either comparing the automated segmentation software within standard neuroimaging analysis packages including SPM, FSL and FreeSurfer or on the optimisation of a single application (Clarkson et al. 2011; Eggert et al. 2012; Fellhauer et al. 2015; Gronenschild et al. 2012; Iscan et al. 2015; Katuwal et al. 2016; Kazemi & Noorizadeh 2014; Klauschen et al. 2009; McCarthy et al. 2015). In short, using these methods on phantom data has shown that both SPM8 and FSL FAST (version 4.1) are reliable and accurate, whereas FreeSurfer (version 4.5) appears to be highly reliable but not necessarily accurate for measuring GM volume (Eggert et al. 2012), whereas SPM 5 and FSL were recommended for GM sensitivity in phantom and control data (Klauschen et al. 2009).

Typically, GM segmentation tools have been developed and optimised for use on healthy brains (Irimia et al. 2012), and therefore may not show the same level of accuracy and reliability when used in clinical cohorts. Given the challenges associated with performing MRI scans in clinical groups, such as the effects of movement, the scans from a clinical research study may be of lower quality due to increased movement and reduced tissue contrast (Kong et al. 2012). Klauschen et al. (2009), for example, found that GM volume is often underestimated in poor quality images with poor contrast and noise, and overestimated in good quality images, indicating possible bias towards reduced GM in patient populations. Furthermore, greater anatomical variability is likely due to pathology in clinical cohorts leading to poor segmentation performance, particularly if software was not designed to cope with pathological abnormalities (Irimia et al. 2012). These key factors are thought to lead to inconsistent findings within clinical neuroimaging studies (Ashburner et al. 2016; Katuwal et al. 2016).

There is evidence that some GM segmentation tools are sensitive to volumetric change in clinical populations, despite their validation in non-clinical populations. SPM8, SPM12, FSL 4.1.9, FreeSurfer 5.1.0 are all sensitive to disease-related change in Alzheimer's disease (Fellhauer et al. 2015). SPM shows the best results for scans with increasing noise (Fellhauer et al. 2015), but performs with greater accuracy for CGM than subcortical GM as shown in Multiple Sclerosis patients (Derakhshan et al. 2010). However, SPM, FSL and FreeSurfer have all shown significant bias in GM measurements when comparing participants with Autism Spectrum Disorder to control participants, indicating that group differences may be due to segmentation related effects rather than true differences (Katuwal et al., 2016); and SPM has been demonstrated to overestimate group differences in healthy elderly participants with atypical anatomy, although this was in a VBM study (Callaert et al. 2014). Although these studies provide some explanation for the difficulty experienced with replication in sMRI studies, especially in clinical participants, these tools are regularly applied to patient cohorts without optimisation for unique brain pathology.

In the following project a number of software tools used to process GM volume were examined to find the most reliable and sensitive means of detecting CGM change in HD participants from the TRACK-HD cohort. Initially, the methods were validated on the BrainWeb dataset, a set of 20 phantom scans with corresponding 'ground truth' segmentations (Aubert-Broche et al. 2006). Following this, 100 TRACK-HD participants were processed to examine performance of the tools in both controls, pre-HD and manifest HD gene carriers. In the baseline 2008 TRACK-HD visit, all 100 participants had back-to-back MRI scans collected in the same scanning session. These back-to-back scans were used to determine the reliability of all tools. The 2011 follow-up scans allowed a comparison of sensitivity to change in the segmentation tools. Both qualitative and quantitative analyses were conducted to ensure that all aspects of performance were examined. Following the analysis, a number of methods have been selected to apply to the clinical cohort in subsequent chapters of this research.

3.1. Aims

This project aimed to investigate the most accurate, reliable and sensitive methods of measuring cross-sectional CGM volume and longitudinal CGM volume change in HD. There was a focus on comparing methods that can be applied by a novice user rather than someone with advanced technical skills. It was deemed important that the results represent common clinical users of neuroimaging software in order to be replicable to the wider community, rather than users who can create their own software or perform detailed optimisation of the available software. In this way, the study aims to encourage use of more consistent methods within the

field and thus allow for better reproducibility of neuroimaging findings, whilst still being appropriate for further projects in this thesis. Total, cortical and lobular grey matter metrics were examined.

3.2. Methods

3.2.1. Segmentation techniques considered

A number of methods are available for delineating (segmenting) GM volume and many of these were considered for inclusion in the study with a final selection made after initial review.

3.2.1.1. Manual delineation via MIDAS

Manual delineation of CGM was initially proposed for comparison. It would be performed using MIDAS, previously described in Chapter 2 (Freeborough et al. 1997). While semi-automated whole-brain segmentation is often considered a 'gold standard' of volumetric segmentation, there is currently no validated technique for measuring only the CGM via MIDAS. Some initial testing was performed to ascertain the feasibility of developing a manual measure. Previously segmented whole-brain regions, which already have the surface between the CSF and GM delineated, were loaded. Thresholds, based on mean brain intensity, were applied to these regions to try to exclude WM. It was possible to achieve a GM region via this method, however different scanner types resulted in different contrast between the GM and WM and thresholds did not perform consistently for all brains. In addition, this technique was labour intensive with each slice requiring checking and manual editing due to the convolutions of the cortex, particularly where there was poor grey/white matter contrast. Due to the considerable time investment required for development, validation and application of a reliable technique using MIDAS, it was ruled out for inclusion in this study. In addition, since MIDAS is not currently a widely available software any measure developed would not be easily reproducible for other research groups. Manual GM measures were excluded from the analysis from this point onwards.

3.2.1.2. FAST (FSL)

As mentioned in Chapter 2, FSLs FAST segmentation procedure is a frequently used software tool (Zhang et al. 2001). FSL is freely available, commonly used in neuroimaging research and easy to implement.

3.2.1.3. SPM

SPM is freely available, although it does require Matlab (which is not freely available) in order to run. It is widely used within neuroimaging research and has been used in a number of HD studies previously (Tabrizi et al. 2009; Hobbs, Henley, et al. 2010; Tabrizi et al. 2011)

3.2.1.4. BRAINS

While BRAINS has been used in some previous HD studies, particularly to quantify subcortical volume of the caudate and putamen, (Nopoulos et al. 2011; Paulsen et al. 2006), and was used in the TRACK-HD and TrackOn studies to measure volume of the putamen (Tabrizi et al. 2009; Tabrizi et al. 2011; Tabrizi et al. 2012; Tabrizi et al. 2013), it is not widely used, is not easy to use, and is somewhat out-dated. Discussion with the software developer responsible for implementing BRAINS in the TRACK-HD study prompted the decision to exclude it from the current study, after the suggestion that it has been surpassed by software that is more capable of accurately measuring GM volume.

3.2.1.5. FreeSurfer

FreeSurfer is commonly used, and has been applied within HD research a number of times, although most analyses are used to perform CT analysis. There are standard pipelines available, and FreeSurfer is easy to implement and freely available.

3.2.1.6. ANTs

Although ANTs is a more recently developed tool and has not been used in HD studies previously, it is freely available and becoming more widely cited in neuroimaging literature. It has variable options for processing datasets, with some fully automated streams available. ANTs pipelines are freely available and online support is provided.

3.2.1.7. MALP-EM

MALP-EM is another more recently developed tool that provides a fully automated segmentation, and has been demonstrated as able to cope with severe brain pathologies (Ledig et al. 2015). It has not been used in HD research to date, but it is freely available and has a recommended pipeline.

3.2.1.8. Final Measures

The final methods for this analysis were selected based on ease of access, frequency of use within the literature and usability of the software. The final measures selected for the GM volumetric methodology comparison were: FSL FAST using FSL version 5.0.9, SPM version 8 Unified Segment, SPM version 8 New Segment, SPM version 12 Segment, FreeSurfer version 5.3.0, ANTs version 2.1.0 (final application compiled on 13-October-2015), and MALP-EM version 1.2.

3.2.2. Cohorts

3.2.2.1. BrainWeb

The BrainWeb cohort was used for an initial analysis. The cohort is described in section 2.1.1.

3.2.2.2. TRACK-HD

Following on from the BrainWeb data analysis, the tools were applied to a subset of the TRACK-HD cohort in order to study the performance of these segmentation measures on atypical brains. The TRACK-HD cohort is described in section 2.1.2. Back-to-back scans from the baseline of TRACK-HD were used to examine the consistency of these tools. As there is no GT, importance was placed on visual checking of the output of all processing pipelines to examine accuracy. The same participants were also examined at their 2011 time point to assess longitudinal change.

100 participants from the TRACK-HD 2008 time point were selected for this study. Twenty controls, 20 pre-A, 20 pre-B, 20 HD1 and 20 HD2 participants were used. The participants were initially considered for inclusion based on whether they had back-to-back scans at baseline of TRACK-HD and follow up scans at the 2011 time point. Following this criteria, participants were randomly selected whilst trying to maintain an approximate age matching between the groups. Controls were examined first, followed by HD gene carriers to highlight potential issues that may occur when applying standard segmentation tools to patients.

3.2.3. Segmentation, quality control and volume calculation

Segmentations for FSL, FreeSurfer, ANTs and MALP-EM were run on a computer cluster (Legion@UCL). All SPM segmentation methods the segmentations were run on a Dell

Precision Power 7810 desktop computer. Following any manipulation of the images, scans were visually examined for quality. For this study, a processing step was categorised as a 'fail' if there was gross failure in performing the extraction or segmentation (Figure 3.1), rather than for more minor errors. Gross failures at any stage were initially checked to rule out user error. Processing changes were made to rectify software-based gross errors in two cases. Firstly, delineation of the GM performed by SPM8 Unified Segment may fail if the orientation of the brain deviates noticeably from the standard SPM templates (Figure 3.1). In a few cases adjustments were performed to shift the brain within the field of view to achieve a better match with standard space. Secondly, prior to segmentation using FSL, the Brain Extraction Tool (BET; Smith 2002) was run on all data. Using standard parameters BET failed to extract brains satisfactorily on the TRACK-HD cohort (Figure 3.1) and optimised BET parameters were substituted (see section 3.4.2). In addition to a small number of gross errors, all tools showed minor errors in segmentation, as discussed in the results (also shown in Figure 3.4). No manual intervention was made in the case of errors in segmentation because this would introduce increased subjectivity to the comparison. Recommendations for manual intervention are provided in the discussion.

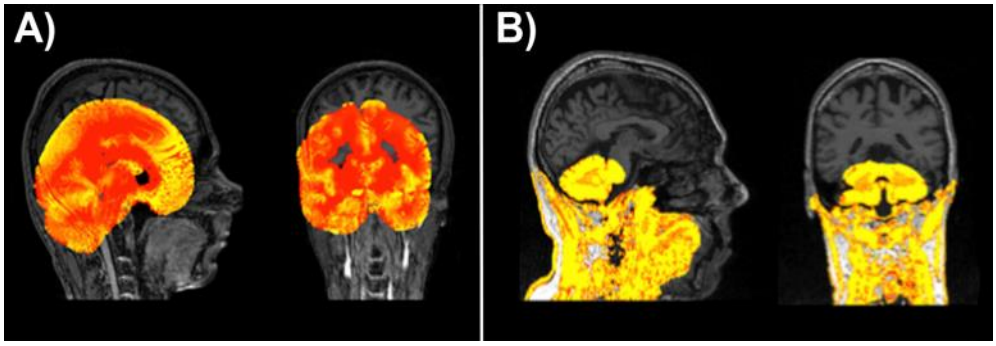


Figure 3.1. (A) An example of a gross failure on a TRACK-HD scan when using SPM8 Unified Segment. (B) An example of a gross failure on a TRACK-HD scan when using FSL BET brain extraction and FAST segmentation procedures.

Partial volume estimates (PVE) were used for all tools except FreeSurfer, which instead of outputting either partial volume or discrete tissue maps outputs text files with regional volumes listed. These are calculated to account for PVE. Due to this, FreeSurfer volumes were calculated differently to other techniques in this analysis (see section 3.2.3.3 for details). Whilst probabilistic segmentation maps were used throughout this analysis as they should more accurately represent brain volume (Tohka 2014) for the first part of the analysis examining total GM volume in BrainWeb data two analyses were run in parallel. The first analysis was run on the discrete volumes and the second on partial volume segmentations.

This was done to allow for some understanding of how PVE and discrete volumes perform. Detailed information on the processing for each programme follows.

3.2.3.1. SPM

All versions of SPM were applied in native space and with default settings for both BrainWeb and TRACK-HD data. SPM only outputs partial volume maps, and so for the first BrainWeb analysis, all grey-matter segmentations were binarised at a threshold of 0.5 via FSLmaths in order to only include voxels that were classed as a majority grey-matter. For the remaining analyses the partial volume GM maps were used.

3.2.3.2. FSL

Prior to FAST being run, the Brain Extraction Tool (BET) was run on all data. For the BrainWeb data this was done with default settings in native space, but as previously mentioned optimised parameters were used for TRACK-HD data. The commands were:

1. London and Paris participants:
   bet <input> <output> -f 0.65 -R -c 104 128 128
2. Leiden participants:
   bet <input> <output> -f 0.3 -R -c 82 120 120
3. Vancouver participants:
   bet <input> <output> -f 0.3 -R -c 82 112 112

Following BET, FAST was run with default settings in native space. FAST outputs both partial volume tissue classes and discrete tissue classes for each image. For the first BrainWeb analysis the discrete GM was used, for all subsequent analyses partial volume maps were used.

3.2.3.3. FreeSurfer

FreeSurfer was run via the default recon-all pipeline, with a flag that specifies that the scans were collected on a 3T scanner. The surface-based stream is mainly used for measuring CT and the volume-based stream is mainly used for calculating volume. Volumetric maps displaying PVE results are not output, meaning that a user cannot simply output the volume of a map. However, the FreeSurfer pipeline includes the automatic calculation of the volume of different regions. It combines the CT and volumetric results during processing and outputs a text file for each participant containing volumetric results. For example, CGM volume is output by FreeSurfer, but it is not calculated by simply measuring the size of the volumetric segmentation, instead it is calculated via a combination of volumetric and surface-based factors (https://surfer.nmr.mgh.harvard.edu/fswiki/MorphometryStats).

This means that it is difficult to replicate the volumes output from the FreeSurfer pipeline via manual extraction from the segmentations, and it is impossible to view the segmentations that contain the exact volume output within the volumetric text files. Based on the recommendations of FreeSurfer developers, the automatically optimised volumes were extracted for each participant and used in the current analysis (https://surfer.nmr.mgh.harvard.edu/fswiki/MorphometryStats). See section 3.2.4 for more detail.

3.2.3.4. ANTs

ANTs requires a template and priors to calculate volumes and thickness. The recommended default pipeline was followed and study-specific templates and priors were created for both datasets in this study (Avants, Tustison, Wu, et al. 2011).

3.2.3.4.1. ANTs BrainWeb processing

First, a template was created using all 20 BrainWeb scans. This was done using the command antsMultivariateTemplateConstruction.sh, which is included with ANTs. All settings were default, except that the options to allow for parallel computation were enabled. Following this, study-specific priors were created from the template. This was done via running the command antsCorticalThickness.sh on the template, using a downloaded template and priors as the required input for the brain mask and priors. The resulting tissue segmentations output from this command were then smoothed, and the CSF segmentation was subtracted from the other 5 segmentations, as recommended by the developers of ANTs (Tustison et al. 2014). The study-specific template, extracted template brain and the corresponding priors could then be used to segment the 20 BrainWeb scans.

ANTs has a segmentation pipeline, Atropos, which can be run in isolation but is also integrated into the CT pipeline along with brain extraction. After preliminary testing it was decided that the CT pipeline would be used for the BrainWeb data as segmentation was not successful using the Atropos pipeline on BrainWeb data. When using the Atropos pipeline, ANTs was unable to accurately distinguish between GM and WM on the BrainWeb scans and resulted in the failure of processing. Instead, the CT pipeline was run with default settings and the template-specific brain mask and priors were included. From the CT pipeline both partial volume tissue classes and regions with each voxel classified as a discrete tissue class are output. For the first analysis the discrete GM was used. For the second analysis, partial volume was used.

3.2.3.4.2. ANTs TRACK-HD processing

For the TRACK-HD processing, a template was created for both of the back-to-back datasets from baseline, and an additional template was created for the final time point using the same participants. The templates and priors were created as described above using a subset of 25 TRACK-HD scans, 5 from each group following the process recommended in (Tustison et al. 2014). Following template creation, ANTs brain extraction was run on each participant using the brain mask from the study-specific template and study-specific priors. Atropos was run on the TRACK-HD data as it was easy to run and segmented the regions successfully. To run Atropos the extracted brain mask for each participant was used with the study-specific priors.

3.2.3.5. MALP-EM

MALP-EM was run using the default settings on both datasets. MALP-EM outputs both discrete and partial volume maps.

3.2.4. Mask selection and registration

A number of the selected segmentation software output combined cortical and subcortical GM segmentations, and it was decided that for the CGM and lobular volumes masks would be overlaid on total GM for each segmentation to output regional volumes rather than using each tools' own regions. While possibly reducing the performance of some tools slightly by using a mask not optimised to that technique, consistency between techniques was ensured this way. One mask for each region was registered to each participant's native space scan and used for all segmentation outputs.

Time was spent selecting both the best registration tools and the best cortical and lobular GM masks to use. The CGM mask had to incorporate total CGM and exclude subcortical GM and cerebellar GM. In addition, the mask needed to be loose enough to capture any regions in the dura or WM that were incorrectly classified as CGM, as often happens in these segmentation techniques. Initially, the Harvard Oxford cortical mask was applied (Desikan et al. 2006). This was extracted in MNI space and registered via FLIRT and FNIRT to native space. The cortical coverage was good and it provided a balance between too loose and too tight. However, the mask included a lot of cerebellar GM. A cerebellum mask was then used to exclude cerebellar GM. Initially, the MNI cerebellar mask was used. After registration this mask was too loose, and overlapped the occipital CGM. Instead, a binary mask of the Buckner tight cerebellar segmentation was used (Buckner et al. 2011). This was downloaded in FreeSurfer-

conformed space, and registered to MNI before being registered to native space for FSL and SPM outputs. It overlaid the cerebellum reasonably well, although in posterior regions it missed some sections of the cerebellum.

It was noted that during registration, the masks were undergoing warping around the exterior surfaces and so following selection of the most appropriate cortical mask and cerebellar exclusion mask, a different registration tool was tested. The ants command, antsRegistrationSyNQuick.sh was used with default settings to convert the standard MNI 1mm brain from MNI space to native space for each participant. The resulting warp was then applied to the Harvard Oxford cortical mask and the cerebellar mask, with much greater alignment seen than when using FLIRT and FNIRT. This registration was also faster to perform and apply. An example of the CGM mask is shown in Figure 3.2.

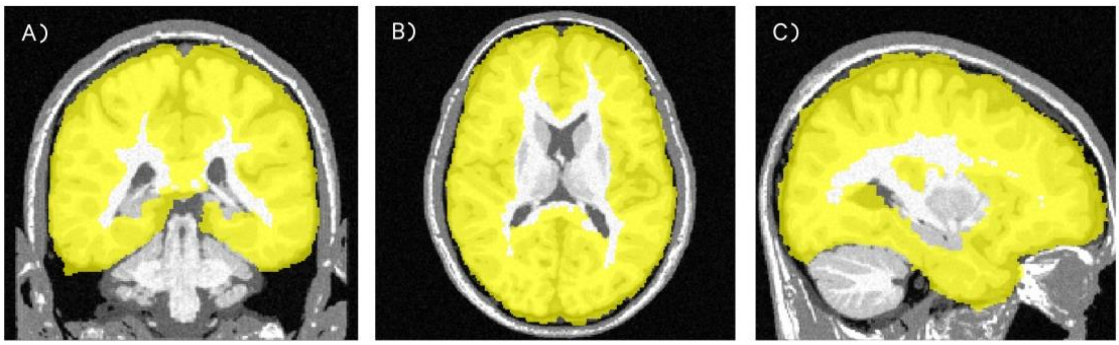


Figure 3.2. The CGM mask for BrainWeb scan 04, overlaid on the native T1 image shown in (A) coronal, (B) axial and (C) sagittal views.

After the initial analysis, it was also decided that a lobular analysis would be performed to provide additional information on the performance of these tools. However, the Harvard Oxford cortical mask does not provide lobular regions. Due to this, it was decided that the MNI GM template included with the FSL package would be applied to output lobular volumes (Mazziotta et al. 2001). This mask was registered to the output images in the same manner as the MNI mask and the frontal, temporal, parietal, occipital and insular regions were extracted. An example of the lobular masks is shown in Figure 3.3.

The only exception to this was FreeSurfer, as mentioned previously, where total GM and CGM volumes were extracted from an automatically created text file. The volumes used in this study were taken from the aseg.stats file created for each participant. The volumes were treated as partial volumes, as it is stated on the FreeSurfer website that the aseg.stats file does control for PVE (https://surfer.nmr.mgh.harvard.edu/fswiki/FsTutorial/QuestionAnswers).

However, the lobular volumes were calculated differently for the BrainWeb and TRACK-HD datasets. FreeSurfer provides additional steps to output lobular volumes, using pre-defined regions of interest. For the BrainWeb study, the default FreeSurfer lobular values were extracted. Based on the results of this analysis, it was decided that for the TRACK-HD analysis the lobular masks used to calculate lobular volume for all other techniques should be used (see section 3.2.4). To do this, the lobular masks for each participant were transformed from native space into each subject's FreeSurfer analysis space (https://surfer.nmr.mgh.harvard.edu/fswiki/FsAnat-to-NativeAnat) and used to extract volumes. As FreeSurfer regions have undergone significant optimisation during development, we would expect that FreeSurfer may have an advantage over the other tools and this should be considered when exploring the results.

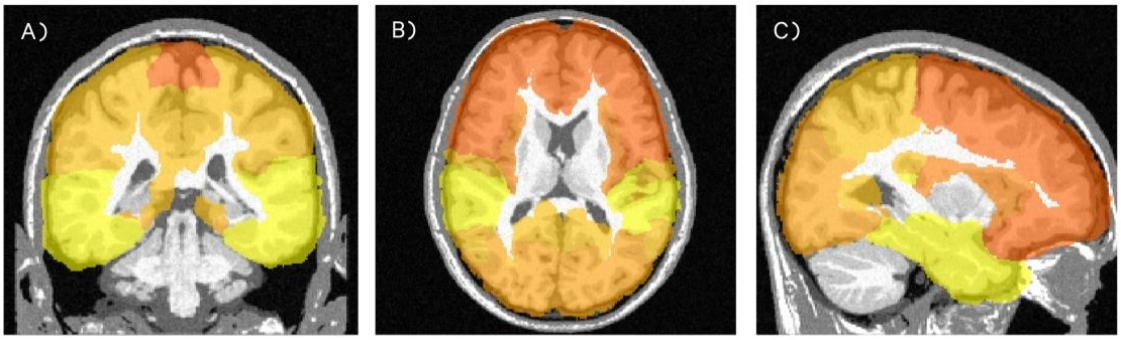


Figure 3.3. The lobular masks for BrainWeb scan 04, overlaid on the native T1 image shown in (A) coronal, (B) axial and (C) sagittal views.

3.2.5. Analysis

A qualitative examination of the data was performed prior to the quantitative analysis. There are various features of these segmentations that can only be understood by examining the volumetric maps. For example, minor but consistent over- or underestimation of GM within automatically- identified anatomical regions may not be detected by quantitative methods, but can be easily recognised by a neuroanatomical expert during visual examination. All scans were examined blinded to group and all errors were noted. From this, trends within each tool were detected. Over- and under-segmentation refers to regions whereby the boundary of the segmented region differed from the visible boundary on the T1. All segmentation tools also showed minor errors as discussed in the results section.

3.2.5.1. BrainWeb

Initially, summary statistics and plots were output to get an overview of the data. Paired t-tests were also conducted to determine whether differences between the GT and segmentation procedures were statistically significant. Scatterplots were produced for a visualisation of the relationship between the volume of the GT and each segmentation procedure. Following this, Bland Altman plots were output along with Pitman's Test of difference in variance to test the agreement between the ground-truth and each segmentation method. Overlap scores (dice coefficient) were also calculated on a voxel-wise basis as discussed in Crum, Camara and Hill (2006). These were calculated for both discrete GM volumes and partial volume maps, but were not calculated for FreeSurfer as voxel-wise maps are not available for the FreeSurfer data.

3.2.5.2. TRACK-HD

Total and CGM volumes were examined in all analyses, and a subset of analyses also examined lobular GM volume. First, summary statistics, including means, ranges and standard deviations for demographic information were calculated and between-group differences derived. Total, cortical and lobular GM mean volumes were calculated for both baseline and the follow-up scan for each participant. Additionally, reliability was tested for the back-to-back baseline scans using intraclass correlation (ICC; Bartko 1966). Mean repeatability for back-to-back scans was also calculated as percent variability error, as shown in Function 1 in section 2.4.1 (Jovicich et al. 2013; Tustison et al. 2014). Spearman's Rho was also used to test the correlation between each set of volumes extracted using each segmentation tool.

To measure the longitudinal sensitivity of each segmentation tool the control group was compared to all HD groups. Follow-up volume of total GM, CGM and lobular volumes were expressed as a percentage of baseline volumes and regression analyses were performed to determine whether there were significant differences in the rate of change between controls and each HD group. All results were adjusted for age, sex and site.

3.3. Results

3.3.1. Qualitative Results

The output for each BrainWeb and TRACK-HD dataset and each tool were visually inspected. Notes were taken for every segmentation examined, and common themes were noted for the different segmentations. This examination was focused on CGM, with subcortical

structures not examined in detail. Table 3.1 provides a description of the performance of each tool on both datasets, and Figure 3.4 contains examples.

3.3.2. BrainWeb Quantitative Results

3.3.2.1. Whole-brain results

Table 3.2 shows summary statistics for discrete and PVE total GM volumes, results of paired t-tests and summary statistics for the dice score. All techniques significantly underestimated grey matter compared to the GT in discrete segmentations (Figure 3.5 A); however the dice overlap values were still high for all techniques, ranging from .890 for ANTs to .920 for FAST.

Figure 3.6 shows the relationship between the discrete GT volumes and discrete volumes output by each segmentation technique, indicating good correspondence for all techniques. In the discrete volumes, bias was detected via Pitman's test of variance in all techniques excluding SPM8 Unified Segment (Figure 3.7). This indicates that the software had a tendency to underestimate GM volume compared to the GT in small brains, and overestimate it in larger brains.

For PVE volumes, all versions of SPM, FAST and FreeSurfer significantly underestimated GM volume whereas ANTs and MALP-EM overestimated GM volume (Figure 3.5 B). The overestimation for ANTs was significant but not for MALP-EM. Again overlap values were high, this time ranging from .882 for FAST to .904 (SPM8 New Segment).

Figure 3.8Figure 3.8 shows the relationship between the discrete GT volumes and discrete volumes output by each segmentation technique. There was again bias in all techniques for the PVE volumes (Figure 3.9).

Table 3.1. A description of the performance of each tool, with most common issues outlined.

Segmentation	BrainWeb Performance	Track-HD Performance
Ground Truth	– In a number of cases dura and skull was included in the volumes. – The GT volumes are somewhat inaccurate, and may over-estimate the actual GM volume in some cases.	--
SPM8 Unified Segment	– SPM8 Unified segment performed well on the BrainWeb scans – Tight around the GM/CSF lateral boundaries – Some WM voxels classed as GM – Good delineation of the sulci	– Poor temporal delineation very common. – Occipital spillage and underestimation of frontal lobes in a number of scans. – 6/400 scans segmented excluded from analysis due to gross failure (1 for 2008 A, 2 for 2008 B non-registered, 2 for 2008 B registered, 1 for 2011).
SPM8 New Segment	– Large number of voxels located within the skull and dura consistently classed as GM, could erroneously inflate the value total volume	– Poor temporal delineation very common – Occipital spillage in a number of scans – Classified voxels in the skull and dura as GM in almost all segmentations – No scans failed segmentation
ANTs Atropos	– GM segmentations accurate overall, with good delineation of the sulci – Some overestimation of GM on WM/CSF boundary	– Variable performance – Brain extraction determined segmentation quality; e.g. large brain mask meant dura included in the segmentation – Frequent overestimation of occipital GM and poorly delineated temporal lobes – No segmentation fails
MALP-EM	– Performed well overall, some GM cut-off in a small number of scans due to tight brain mask	– Fewer issues with overestimation of the occipital lobe – Generally better temporal lobe delineation – CGM underestimated in superior regions in a small number of cases – No segmentation fails
FSL FAST	– Performed well overall, clean boundaries although some segmentations classified skull as GM	– Standard BET provided poor brain extractions on Track-HD data and was re-run with an optimised BET procedure, although results of the optimised BET were still sub-optimal – Often underestimated GM volume, with occasional overestimation due to poor brain extraction – As a result, difficult to characterise GM segmentation – Two scans were rated as a complete fail
FreeSurfer	– Performed very poorly overall, large regions of the frontal, temporal and occipital lobes were regularly underestimated and missing from the segmentation.	– GM tended to be very tight along CSF boundary, with a layer of voxels on the GM/CSF boundary typically excluded from the volume based segmentation – Within FreeSurfer volume is calculated via a combination of the volume and surface- based segmentations and so some of these excluded voxels would be included in the calculation if they are within the cortical surface – Spillage into the temporal CSF and occipital dura regularly seen with some cases classifying skull as GM

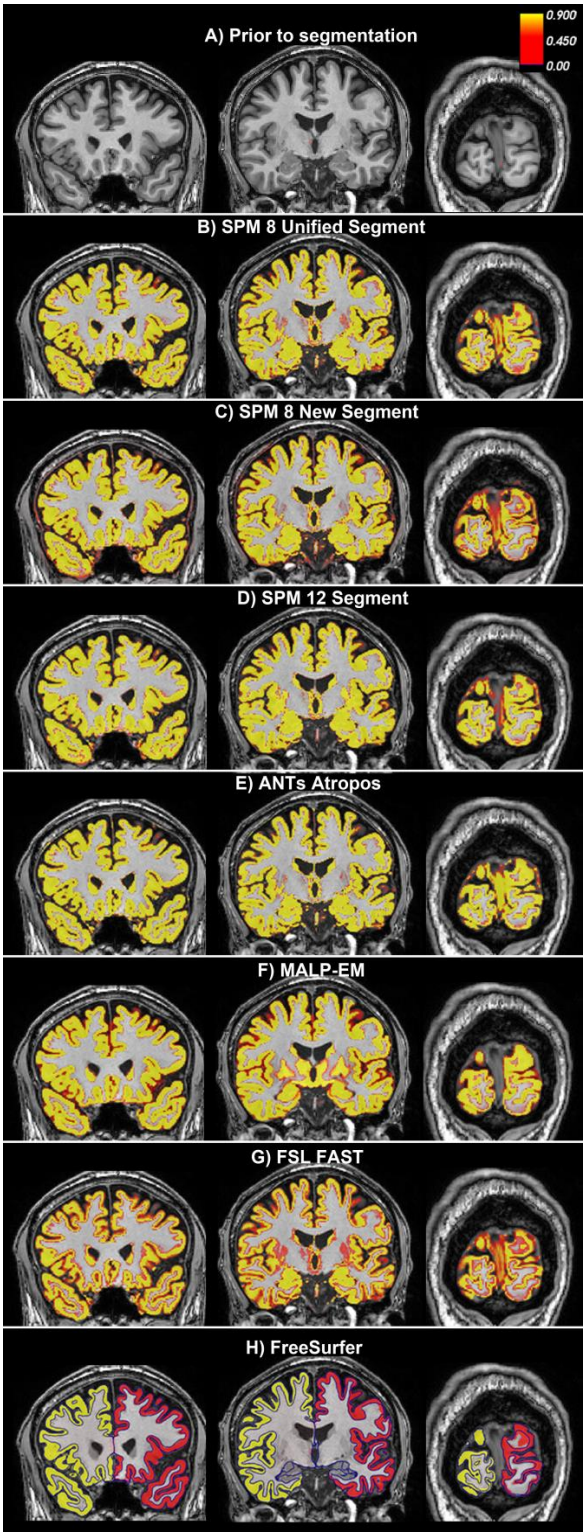


Figure 3.4. Examples of the GM output from each tool overlaid on one participant from the TRACK-HD study. The figure shows three coronal views, with Figure (A) showing the same slices with no segmentation. The first slice shows the frontal and temporal regions, the second slice is towards the middle of the brain, and the last slice shows the occipital lobe. All figures show default probabilistic segmentation maps for each software except for FreeSurfer, which shows volumetric and surface based regions. For the probabilistic segmentation maps, the brighter the yellow within a voxel, the more likely that the voxel contains GM.

Table 3.2. Discrete and PVE total GM volumes for the Ground Truth and all segmentation techniques, paired t-tests showing significant differences between the total GM Ground Truth volume and all segmentation techniques, for discrete and PVE volumes and summary statistics for the overlap metrics between Ground Truth and all techniques for both discrete and PVE volumes.

	Discrete Volumes (ml) Mean (st Dev) Min-Max	Discrete Volumes Paired t-tests	Discrete Volumes Overlap Mean (st Dev) Min-Max	PVE Volumes Mean (st Dev) Min-Max	PVE Volumes Paired t-tests	PVE Volumes Overlap Mean (st Dev) Min-Max
Ground Truth	1123.26 (50.58) 1028.46-1249.82	-		968.54 (54.93) 875.71-1093.04	-	
SPM8 Unified Segment	950.54 (42.65) 864.73-1030.22	-172.713 t = -30.66 (19) p = .0000	0.904 (0.014) 0.867-0.926	947.90 (42.06) 861.36-1023.02	-20.64 t = -3.21 (19) p = .0046	0.903 (0.010) 0.880-0.914
SPM8 New Segment	954.04 (31.84) 887.93-1016.01	-169.22 t = -29.87 (19) p = .0000	0.902 (0.010) 0.873-0.920	950.53 (28.10) 889.18-1007.49	-18.01 t = -2.36 (19) p = .0292	0.904 (0.007) 0.888-0.914
SPM12 Segment	945.52 (38.18) 863.20-1020.92	-177.73 t = -35.35 (19) p = .0000	0.896 (0.013) 0.861-0.919	941.67 (37.81) 857.29-1012.36	-26.88 t = -4.38 (19) p = .0003	0.900 (0.008) 0.885-0.910
ANTs	1031.50 (34.82) 949.57-1096.22	-91.75 t = -16.39 (19) p = .0000	0.890 (0.016) 0.855-0.919	1041.76 (31.21) 967.25-1091.67	73.22 t = 8.94 (19) p = . 0000	0.898 (0.008) 0.884-0.913
MALP-EM	949.32 (368) 862.65-995.46	-173.94 t = -25.42 (19) p = .0000	0.891 (0.011) 0.864-0.912	984.84 (36.45) 893.97-1038.23	16.29 t = 1.87 (19) p = .0777	0.886 (0.010) 0.859-0.899
FAST	906.10 (41.99) 832.49-1002.41	-217.15 t = -56.59 (19) p = .0000	0.920 (0.012) 0.889-0.940	957.07 (47.80) 876.51-1057.84	-11.47 t = -3.72 (19) p = .0015	0.882 (0.008) 0.867-0.897
FreeSurfer	-	-	-	919.91 (40.77) 851.88-1014.43	-48.63 t = -9.64 (19) p = .0000	-

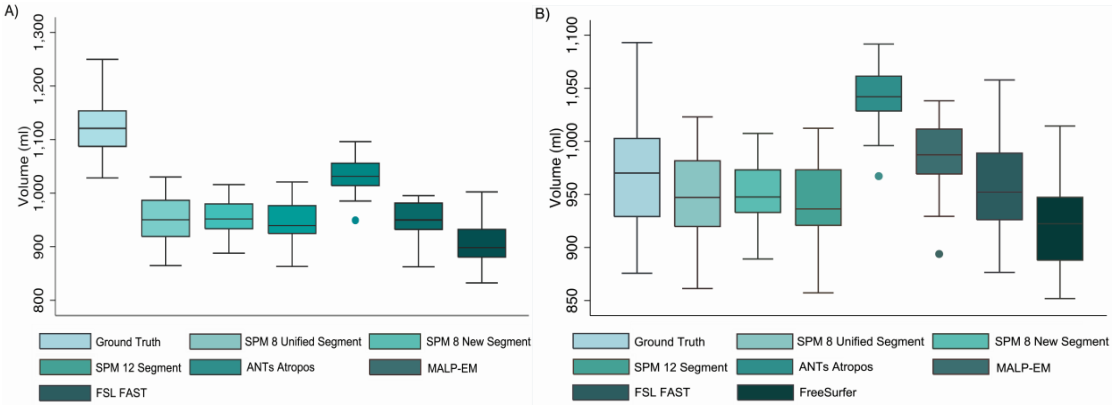


Figure 3.5. (A) Boxplots showing total GM discrete volume for BrainWeb Ground Truth vs. segmentation methods (B) Boxplots showing partial volumes estimates for total GM volume for BrainWeb Ground Truth vs. segmentation methods.

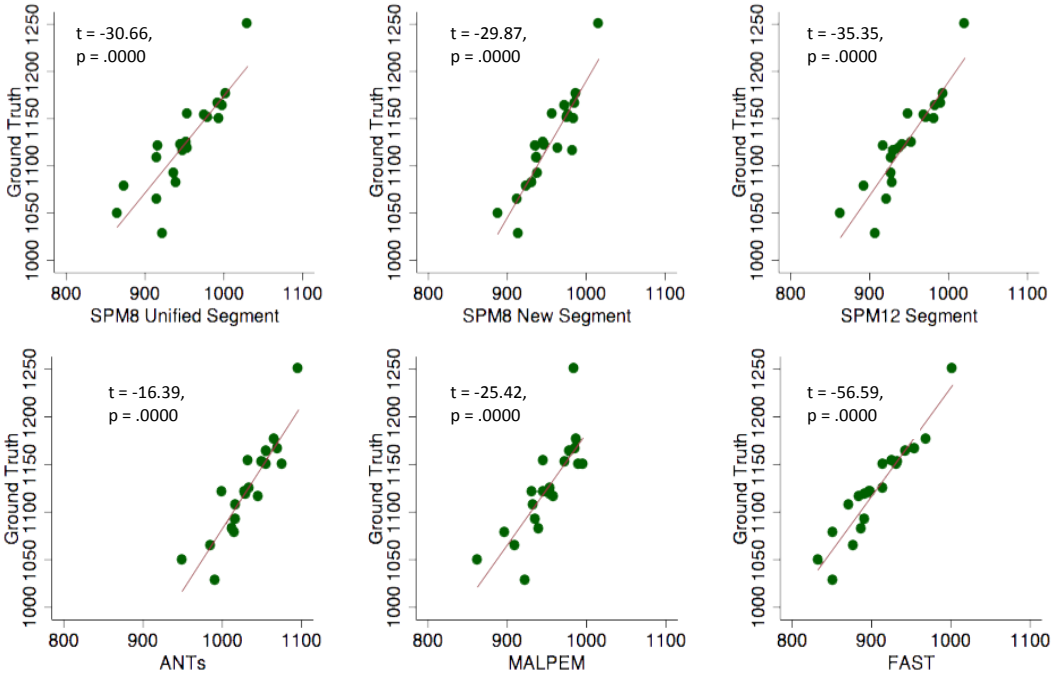
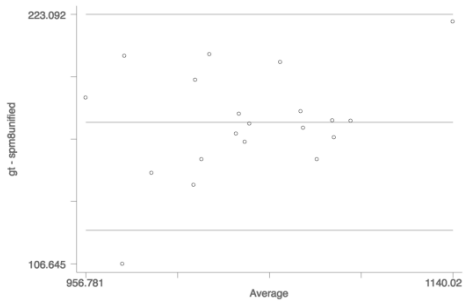
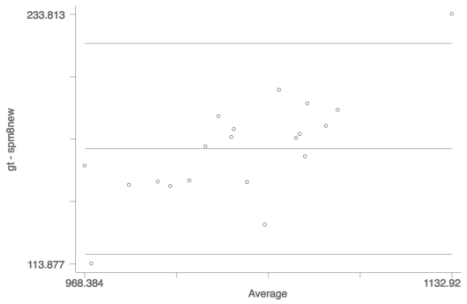


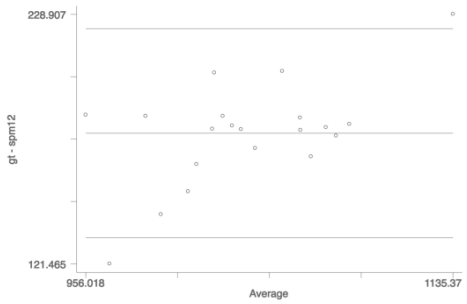
Figure 3.6. Scatterplots showing the relationship between Ground Truth total discrete GM volume and discrete total GM volume for all segmentation techniques. Associated t-statistics and p values are shown for the results of paired t-tests comparing ground truth values and segmented values.



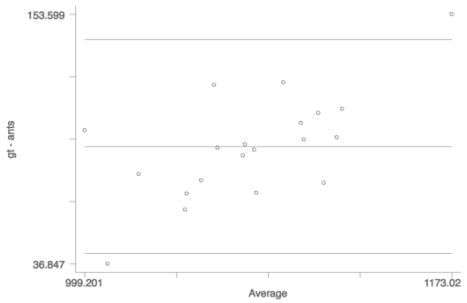
SPM8 Unified Segment
r = 0.325, n = 20, p = 0.161



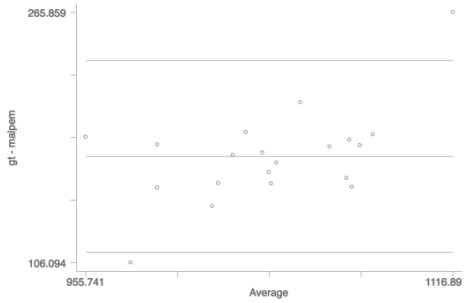
SPM8 New Segment
r = 0.756, n = 20, p = 0.000



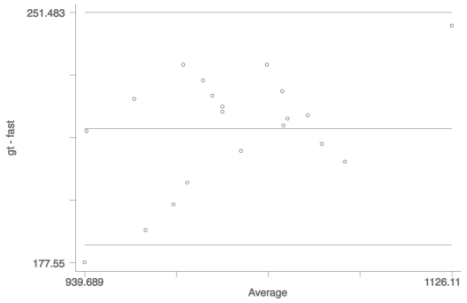
SPM12 Segment
r = 0.564, n = 20, p = 0.010



ANTs
r = 0.646, n = 20, p = 0.002



MALP-EM
r = 0.546, n = 20, p = 0.013



FAST
r = 0.507, n = 20, p = 0.023

Figure 3.7. Bland Altman plots and Pitman's test of variance for the relationship between Ground Truth total discrete GM volume and discrete total GM volume for all segmentation techniques.

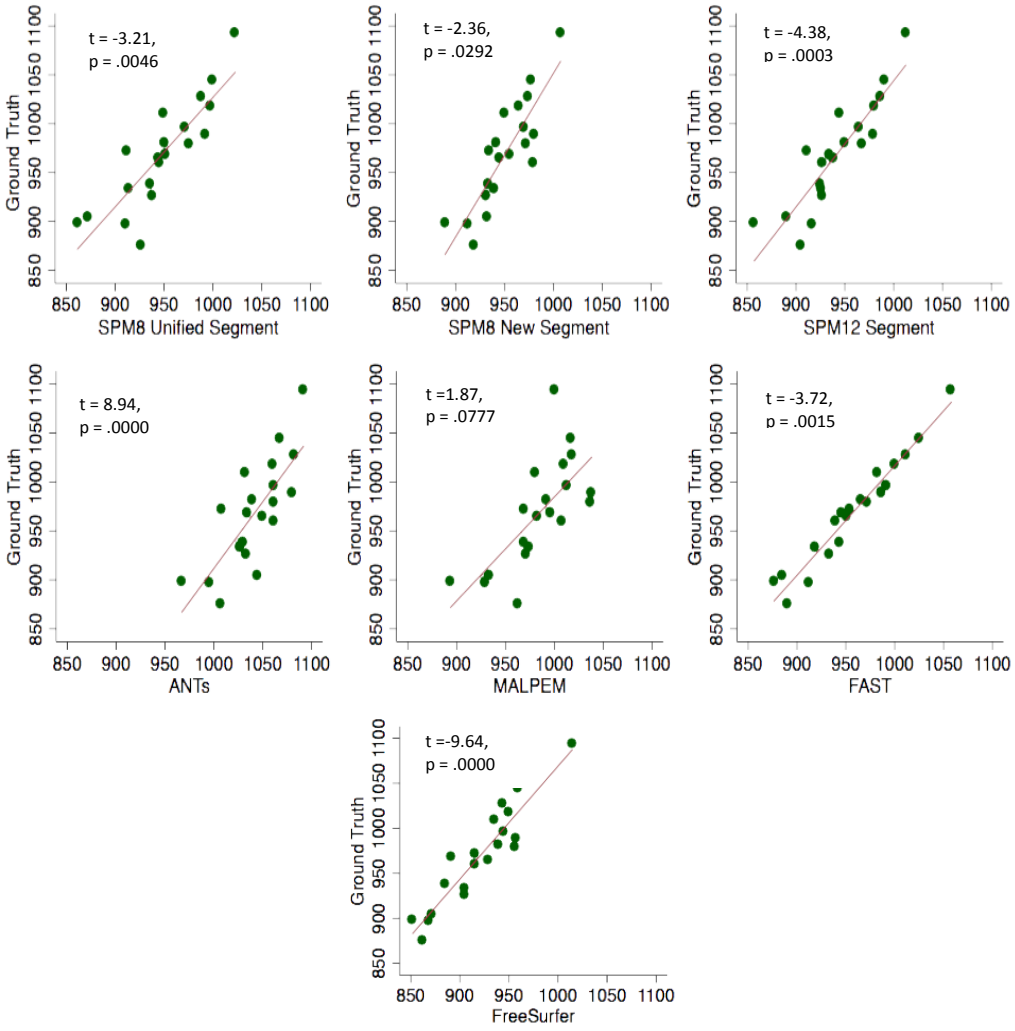
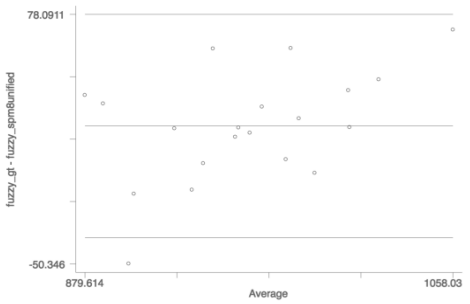
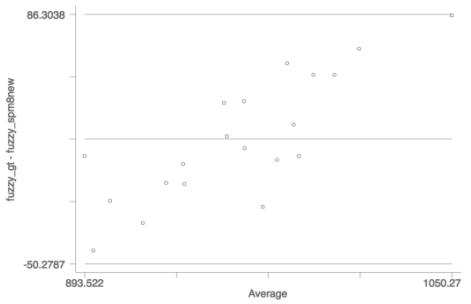


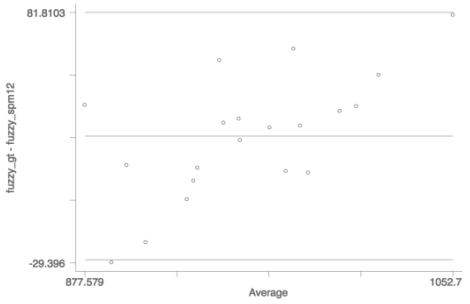
Figure 3.8. Scatterplots showing the relationship between Ground Truth total PVE GM volumes and PVE total GM volume for all segmentation techniques. Associated t-statistics and p values are shown for the results of paired t-tests comparing ground truth values and segmented values.



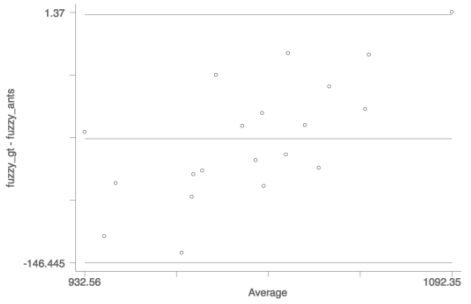
SPM8 Unified Segment
r = 0.465, n = 20, p = 0.039



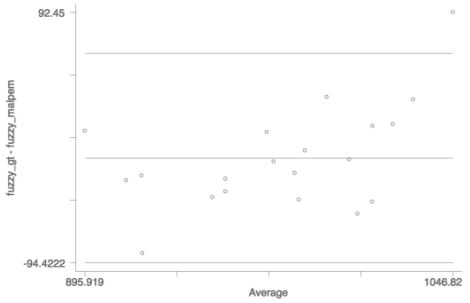
SPM8 New Segment
r = 0.812, n = 20, p = 0.000



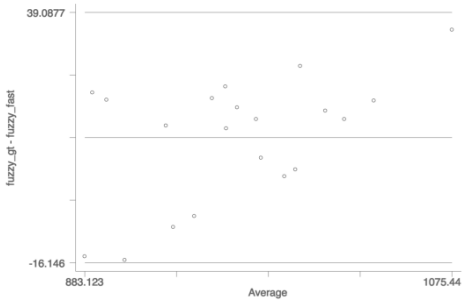
SPM12 Segment
r = 0.641, n = 20, p = 0.002



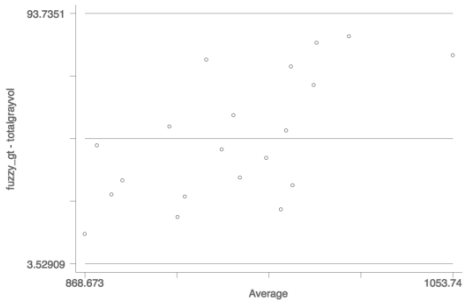
ANTs
r = 0.685, n = 20, p = 0.001



MALP-EM
r = 0.511, n = 20, p = 0.021



FAST
r = 0.520, n = 20, p = 0.019



FreeSurfer
r = 0.639, n = 20, p = 0.002

Figure 3.9. Bland Altman plots and Pitman's test of variance for the relationship between Ground Truth total PVE GM volume and PVE total GM volume for all segmentation techniques.

3.3.2.2. Cortical GM Results

All segmentation methods except ANTs underestimated CGM, as shown in Table 3.3 and Figure 3.10. The difference between the GT mean volume and the mean volumes for each technique were significant for SPM8 Unified Segment, SPM12, ANTs and FAST at $p < .05$. Dice scores measuring the overlap between the GT and segmentations indicated that while all techniques showed substantial overlap with the GT, the lowest average overlap was seen in SPM8 New Segment (0.899), FAST had the largest overlap with total CGM (0.922). Scatterplots showing the relationship between the GT and each technique suggest that FAST also had the closest relationship with the GT (Figure 3.11). Furthermore, Pitman's test of variance indicated bias in measuring CGM for all techniques excluding SPM8 Unified Segment (Figure 3.12). All techniques displaying bias in segmentation overestimated CGM in brains whereby the GT was large and underestimated CGM in brains whereby the GT was small.

Table 3.3. Descriptive statistics for CGM volume in the BrainWeb dataset, paired t-test results showing significant differences between the cortical GT volume and all segmentation pipelines and dice scores for the overlap between each segmentation procedure and the GT for CGM.

	Cortical Volumes (ml) Mean (st Dev) Min-Max	**Cortical Volumes Paired t-tests**	**Cortical Volumes Overlap Mean (st Dev) Min-Max**
Ground Truth	765.47 (44.11) 692.89-867.36	-	-
SPM8 Unified Segment	750.34 (35.68) 676.73-814.57	-7.88 t = -3.36 (19) p =.0033	0.907 (0.013) 0.873-0.928
SPM8 New Segment	755.43 (22.6) 720.71-805.01	-15.13 t = -3.07 (19) p = .0063	0.907 (0.01) 0.882-0.926
SPM12 Segment	746.43 (33.51) 690.81-809.78	-10.04 t = -1.67 (19) p = .1122	0.899 (0.012) 0.868-0.922
ANTs	787.18 (26.97) 740.68-836.33	-19.04 t = -4.16 (19) p = .0005	0.904 (0.013) 0.875-0.928
MALP-EM	761.64 (29.58) 701.22-809.94	21.71 t = 3.90 (19) p = .0010	0.904 (0.01) 0.882-0.922
FSL FAST	757.59 (39.13) 685.55-841.65	-3.83 t = -0.58 (19) p = .5671	0.922 (0.011) 0.895-0.942
FreeSurfer	699.32 (34.07) 639.91-779.9	-66.15 t = -18.53 (19) p = .0000	-

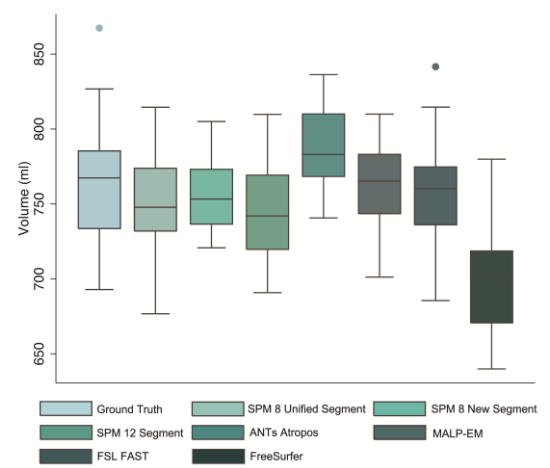


Figure 3.10. Boxplots showing CGM volume for BrainWeb GT vs all segmentation methods

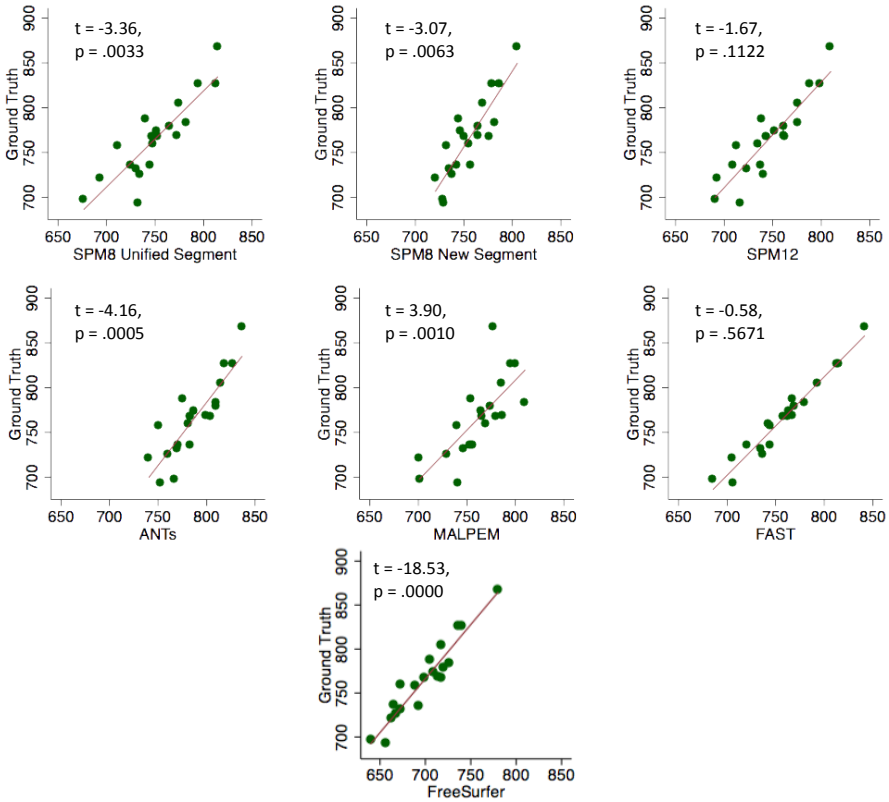
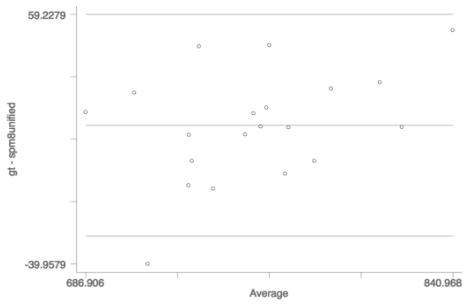
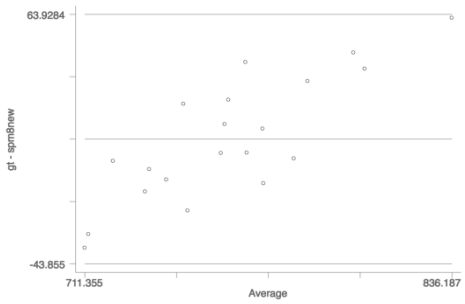


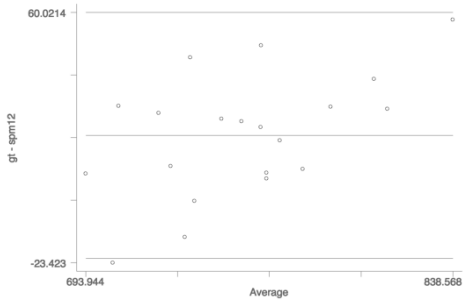
Figure 3.11. Scatterplots showing the relationship between CGM PVE Ground Truth volumes and CGM PVE for all segmentation techniques. Associated t-statistics and p values are shown for the results of paired t-tests comparing ground truth values and segmented values.



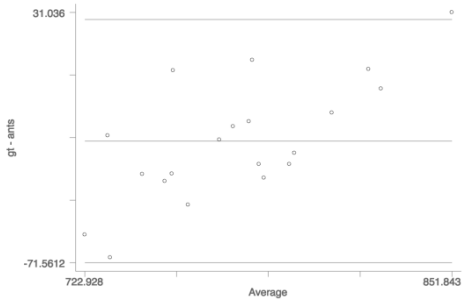
A: SPM8 Unified Segment
r = 0.395, n = 20, p = 0.085



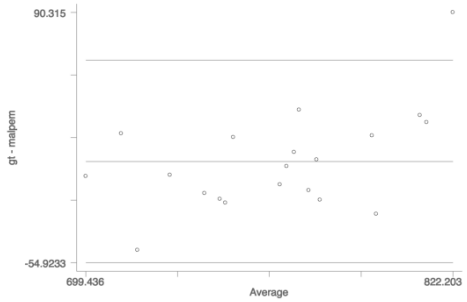
B: SPM8 New Segment
r = 0.823, n = 20, p = 0.000



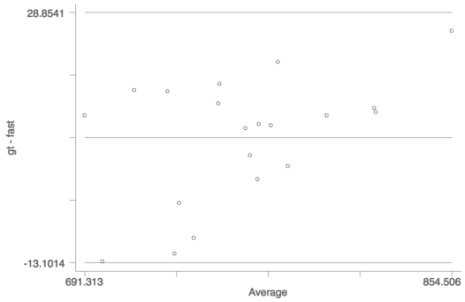
C: SPM12 Segment
r = 0.531, n = 20, p = 0.016



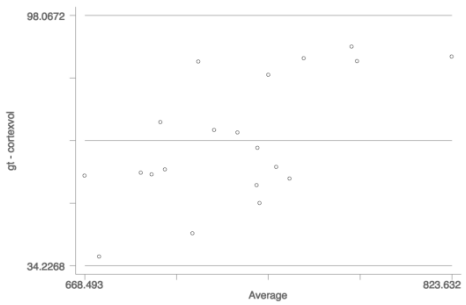
D: ANTs
r = 0.711, n = 20, p = 0.000



E: MALP-EM
r = 0.527, n = 20, p = 0.017



F: FAST
r = 0.477, n = 20, p = 0.033



G: FreeSurfer
r = 0.637, n = 20, p = 0.003

Figure 3.12. Bland-Altman plots showing the agreement between each measure and the Ground Truth for CGM.

3.3.2.3. Lobular GM Results

Volumes, paired t-tests and overlap values are shown in Table 3.4. Mean volumes for the frontal lobe were all underestimated compared to the GT, except for the ANTs segmentation, which overestimated the volume. Frontal volume was significantly underestimated for all versions of SPM, FAST and FreeSurfer, and significantly overestimated for ANTs. The largest difference was between the GT and FreeSurfer. The overlap between GT and segmentations for the frontal lobe was highest for MALP-EM (.934).

Temporal volumes were slightly overestimated by all segmentation techniques, and FreeSurfer greatly overestimated temporal volumes. This overestimation was significant for FreeSurfer, ANTs and FAST. Again the overlap was highest for MALP-EM (dice score = .935), with other techniques also showing high overlap with the GT.

Parietal volume was underestimated in all techniques except MALP-EM and ANTs, with significant underestimation in SPM8 Unified Segment, SPM12, FAST and FreeSurfer, and significant overestimation in ANTs. Overlap values ranged from 0.898 for SPM8 New Segment to .936 for MALP-EM.

Occipital lobes were all underestimated compared to the GT, again with the exception of ANTs, which slightly overestimated the volumes. FreeSurfer, again, greatly underestimated occipital lobe volumes. The occipital lobe differences were significant for all techniques except SPM8 New Segment. MALP-EM showed the highest overlap with the GT.

Insula lobe volumes were overestimated by all techniques, except FreeSurfer, with significant overestimation for SPM8 New Segment, ANTs, MALP-EM and FAST and significant underestimation for FreeSurfer. The overlap with GT was highest for MALP-EM, but again was high for all techniques.

Table 3.4. Descriptive statistics for GM lobes in the BrainWeb dataset, paired t-test results showing significant differences between the Ground Truth and all segmentation pipelines, and dice scores for the overlap between each segmentation tool and the Ground Truth.

		Ground Truth	SPM8 Unified Segment	SPM8 New Segment	SPM12 Segment	ANTs	MALP-EM	FAST	FreeSurfer
Frontal Lobe	**Volume Mean (st Dev) Min-Max**	298.45 (20.68) 264.21-351.21	289.04 (17.08) 257.27-322.37	291.63 (13.26) 270.77-321.49	287.72 (16.06) 259.19-320.78	303.52 (14.26) 278.54-330.62	295.11 (15.25) 263.34-321.68	293.84 (18.43) 262.78-335.59	268.52 (14.02) 249.97-302.92
	Paired t-tests	-	-9.41 t = -5.06 (19) p = .0001	-6.82 t = -3.07 (19) p = .0063	-10.73 t = -6.01 (19) p = .0000	5.07 t = 2.29 (19) p = .0339	-3.34 t = -1.16 (19) p = .2613	-4.61 t = -4.98 (19) p = .0001	-29.93 t = -14.57 (19) p = 0.000
	Overlap Mean (st Dev) Min-Max	-	0.898 (0.020) 0.852-0.921	0.898 (0.015) 0.874-0.924	0.905 (0.020) 0.858-0.941	0.905 (0.015) 0.876-0.931	0.934 (0.012) 0.907-0.955	0.923 (0.012) 0.895-0.942	-
Temporal Lobe	**Volume Mean (st Dev) Min-Max**	99.67 (8.68) 84.49-116.62	101.69 (8.43) 87.28-123.55	99.98 (7.42) 89.14-122.10	100.31 (8.74) 84.57-124.10	107.73 (7.02) 95.29-126.47	101.48 (7.49) 89.02-119.99	100.90 (9.08) 85.06-123.82	155.69 (7.73) 144.45-169.69
	Paired t-tests	-	2.02 t = 1.93 (19) p = .0689	.30 t = .28 (19) p = .7824	.64 t = .71 (19) p = .4868	8.05 t = 7.45 (19) p = . 0000	1.81 t = 1.66 (19) p = .1130	1.23 t = 2.23 (19) p = .0378	56.02 t = 27.74 (19) p = 0.000
	Overlap Mean (st Dev) Min-Max	-	0.898 (0.018) 0.874-0.926	0.897 (0.019) 0.844-0.930	0.906 (0.020) 0.859-0.936	0.906 (0.013) 0.886-0.936	0.935 (0.011) 0.912-0.955	0.923 (0.012) 0.901-0.948	-
Parietal Lobe	**Volume Mean (st Dev) Min-Max**	172.18 (13.34) 150.52-203.15	167.93 (10.97) 145.56-192.22	170.11 (8.67) 154.69-187.40	166.90 (10.23) 149.67-188.27	176.82 (8.70) 163.33-195.05	175.35 (9.28) 157.81-194.78	169.39 (12.03) 147.63-196.49	161.66 (11.93) 143.81-187.48
	Paired t-tests		-4.25 t = -3.26 (19) p = .0041	-2.07 t = -1.28 (19) p = .2175	-5.28 t = -4.26 (19) p = .0004	4.64 t = 3.03 (19) p = . 0070	3.17 t = 1.98 (19) p = .0627	-2.79 t = -4.48 (19) p = .0003	-10.51 t = -9.97 (19) p = .0000
	Overlap Mean (st Dev) Min-Max		0.900 (0.017) 0.869-0.935	0.898 (0.021) 0.843-0.935	0.907 (0.021) 0.851-0.952	0.907 (0.015) 0.871-0.942	0.936 (0.013) 0.900-0.964	0.924 (0.011) 0.892-0.946	-
Occipital Lobe	**Volume Mean (st Dev) Min-Max**	175.25 (9.52) 154.19-193.17	172.25 (8.27) 156.88-190.23	174.78 (7.38) 157.46-186.79	171.93 (8.10) 155.88-188.12	178.15 (7.95) 162.44-192.53	170.73 (7.44) 154.06-183.17	173.70 (8.42) 156.19-190.75	65.08 (5.60) 55.59-74.11
	Paired t-tests		-3.00 t = -3.50 (19) p = .0024	-.47 t = -0.41 (19) p = .6828	-3.32 t = -4.22 (19) p = .0005	2.90 t = 3.24 (19) p = . 0043	-4.52 t = -3.85 (19) p = .0011	-1.55 t = -3.64 (19) p = .0017	-110.17 t = -54.95 (19) p = 0.000
	Overlap Mean (st Dev) Min-Max		0.896 (0.024) 0.841-0.929	0.899 (0.019) 0.855-0.927	0.907 (0.020) 0.877-0.942	0.904 (0.015) 0.881-0.928	0.935 (0.009) 0.916-0.953	0.923 (0.014) 0.897-0.956	-
Insula	**Volume Mean (st Dev) Min-Max**	21.27 (1.59) 18.17-25.39	21.29 (1.57) 17.91-25.13	21.60 (1.53) 18.28-25.06	21.30 (1.49) 18.02-25.11	22.12 (1.53) 18.90-25.80	21.73 (1.57) 18.37-25.24	21.56 (1.59) 18.34-25.70	20.18 (1.09) 17.99-22.17
	Paired t-tests		.03 t = 0.38 (19) p = .7166	.33 t = 3.15 (19) p = .0053	.03 t = 0.41 (19) p = .6830	.85 t = 9.62 (19) p = . 0000	.47 t = 4.47 (19) p = .0003	.30 t = 7.65 (19) p = .0000	-1.09 t = -3.42 (19) p = .0000
	Overlap Mean (st Dev) Min-Max		0.896 (0.021) 0.846-0.924	0.898 (0.017) 0.851-0.921	0.906 (0.022) 0.858-0.942	0.906 (0.013) 0.878-0.928	0.932 (0.012) 0.908-0.949	0.923 (0.015) 0.892-0.956	-

3.3.3. TRACK-HD quantitative results

Participant demographics are given in Table 3.5. Age for the two HD groups was slightly higher than for controls and the pre-HD groups, but this difference was not significant. CAG was significantly higher in PreHD-B, HD1 and HD2 than in PreHD-A, and significantly higher for HD1 and HD2 than for PreHD-B. As would be expected, DBS was also increased in the HD groups. DBS in PreHD-B, HD1 and HD2 were all significantly higher than PreHD-A, and burden in HD1 and HD2 was significantly higher than PreHD-B.

Table 3.5. Demographics for the participants included in the TRACK-HD analysis.

	Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**
Age	48.32 (9.28) 30.73-62.97	48.48 (6.70) 37.27-59.41	47.74 (7.72) 38.11-64.13	49.10 (8.19) 31.11-59.63	50.14 (8.94) 33.26-62.41
Sex	Females N=13	Females N=10	Females N=12	Females N=9	Females N=8
Education	4.05 (1.28) 2-6	4.30 (2.27) 2-6	4.1 (2.03) 2-5	4.2 (1.36) 2-6	3.55 (2.32) 2-6
CAG	N/A	41 (1.21) 39-43	42.35 (1.27) 40-44	43.35 (1.90) 40-47	43.75 (2.45) 41-52
DBS	N/A	259.80 (29.50) 171-290.75	318.43 (23.99) 267.6-356.03	372.35 (52.05) 264.75-472.91	399.15 (70.31) 287.31-548.74

3.3.3.1. Volumetric measures

Volumes of the total, cortical and lobular GM were extracted for control participants and participants at different stages of HD. Total, CGM and lobular volumes are described for both baseline and follow-up time points for each segmentation tool (Table 3.6), and Figures 3.13 and 3.14 show volumes for total and cortical GM. For all techniques, both total and CGM volumes were lower in participants with more advanced disease stage.

For lobular volumes, all tools showed more discrepancies in raw volume (Table 3.6). For the frontal lobe all techniques estimated higher GM volume in PreHD-A participants than in controls, with some techniques also estimating higher GM volume in PreHD-B compared to controls. Frontal lobe volume was reduced in HD1 and HD2 groups compared to control and pre-HD groups. Temporal lobe volumes were also higher in PreHD-A than in controls for all segmentation tools, with both SPM12 and MALP-EM estimating higher temporal volume in all groups compared to controls. Other tools showed slightly lower volumes in PreHD-B, HD1 and HD2 than in controls. Parietal lobe volumes showed more uniform volume differences between each technique, with all techniques except for FreeSurfer measuring higher parietal lobe volumes in PreHD-A, PreHD-B and HD1 compared to controls, and lower volume in HD2

compared to controls. FreeSurfer showed lower volumes in all groups from PreHD-B when compared to controls. For the occipital lobe, the results were variable for each technique, with most techniques showing higher volumes in PreHD-A than in controls, and slightly reduced volume with increasing disease progression. Only FreeSurfer showed large reductions in volume between each group. Finally, insula volume was largest in PreHD-B for all techniques except FreeSurfer, with between-group differences appearing minimal for most tools.

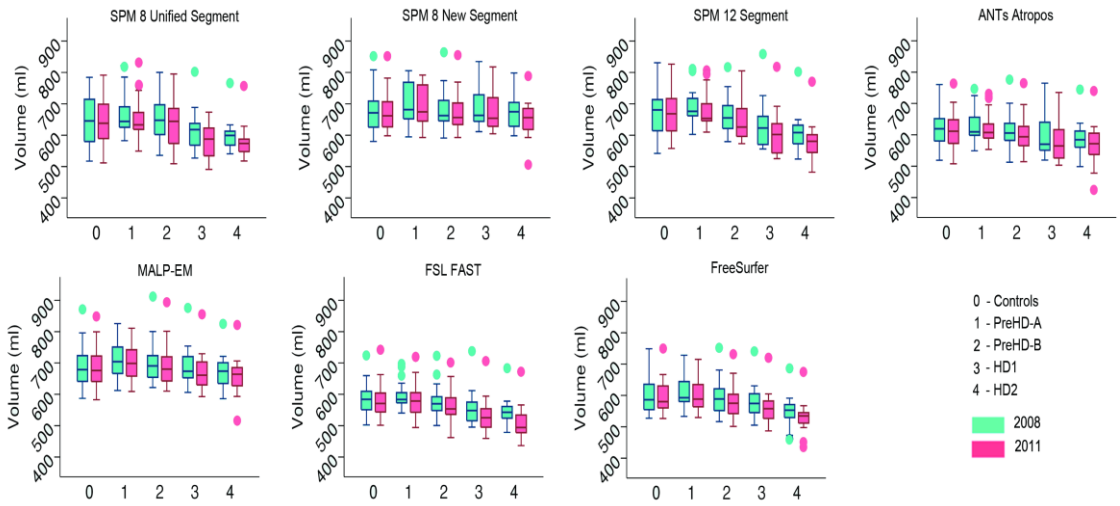


Figure 3.13. Box plots showing total grey matter volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median, and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

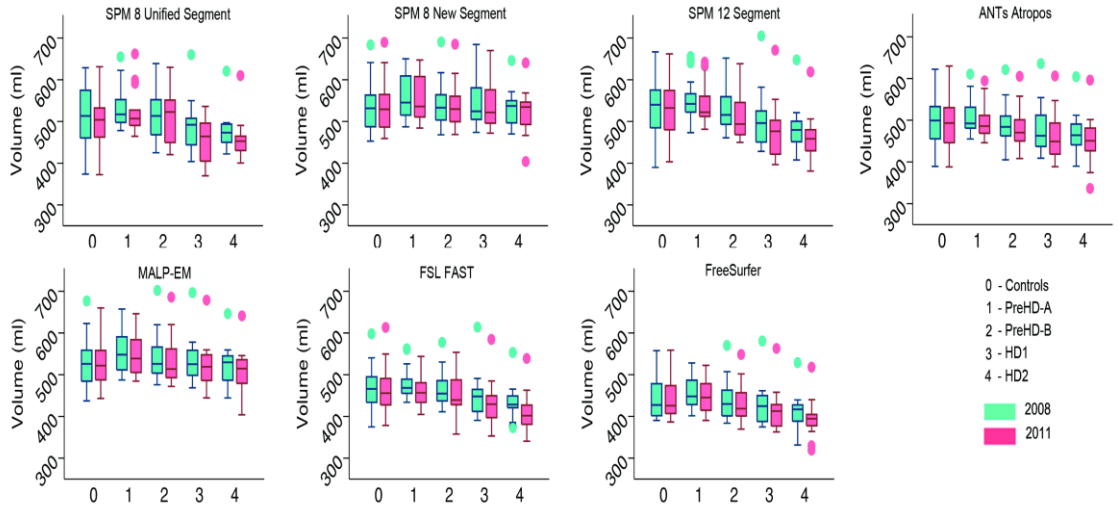


Figure 3.14 Box plots showing cortical grey matter volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median, and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

Table 3.6. GM Volumes for all regions, groups, techniques and time points.

		Controls (N=20)	PreHD-A (N=20)	PreHD-B (N=20)	HD1 (N=20)	HD2 (N=20)
			Total GM			
SPM8 Unified	**2008 Scan A**	644.72 (77.40)	667.41 (64.63)	650.36 (72.40)	617.39 (60.66)	598.58 (48.13)
		517.51-784.50	582.46-818.97	535.75-800.05	527.10-802.07	540.66-766.59
	2008 Scan B	642.71 (75.91)	668.16 (69.54)	651.36 (69.07)	616.83 (62.97)	597.25 (48.72)
		529.61-789.18	582.38-832.93	556.09-804.35	522.50-818.28	533.81-770.85
	2011	642.88 (77.28)	653.41 (67.68)	637.43 (81.73)	583.76 (53.99)	577.33 (51.22)
		511.55-791.30	549.42-831.86	508.94-795.31	491.05-673.19	517.65-757.70
SPM8 New Segment	**2008 Scan A**	678.94 (71.25)	703.62 (66.95)	683.88 (63.96)	682.13 (57.10)	673.07 (47.86)
		580.01-852.75	594.52-805.85	591.00-864.42	611.38-834.77	598.17-798.72
	2008 Scan B	679.28 (72.29)	702.94 (67.23)	684.49 (64.82)	682.36 (57.61)	672.53 (47.82)
		580.84-859.30	592.78-810.64	591.12-870.01	610.99-835.37	597.21-797.95
	2011	675.48 (67.24)	692.85 (64.90)	675.12 (63.58)	669.83 (54.54)	653.26 (55.78)
		598.10-851.22	592.63-791.93	592.57-855.63	604.48-817.85	506.65-788.99
SPM12	**2008 Scan A**	671.50 (72.43)	691.43 (60.49)	663.25 (57.93)	629.56 (72.22)	607.31 (58.86)
		542.23-831.21	602.55-811.37	579.39-817.00	556.09-857.89	523.90-801.68
	2008 Scan B	671.98 (72.57)	689.10 (58.37)	661.92 (55.87)	628.62 (71.94)	604.27 (60.21)
		548.55-832.24	599.37-806.38	577.99-820.37	548.74-854.13	522.27-801.91
	2011	665.82 (70.42)	680.32 (59.33)	645.14 (61.23)	601.98 (72.38)	578.07 (61.30)
		557.72-826.72	610.00-806.47	573.01-805.50	525.19-818.94	482.20-771.05
ANTS	**2008 Scan A**	619.58 (58.45)	631.63 (54.56)	610.96 (58.83)	596.48 (64.76)	587.03 (53.28)
		519.03-760.16	549.13-746.23	512.86-776.31	519.36-764.84	498.50-745.08
	2008 Scan B	614.92 (61.17)	624.97 (55.05)	607.81 (57.48)	592.71 (62.51)	581.45 (57.93)
		510.65-764.75	540.81-740.83	515.70-774.35	504.09-764.36	473.80-745.80
	2011	611.35 (63.58)	620.91 (53.98)	601.04 (61.41)	578.62 (64.08)	568.61 (65.08)
		507.70-763.61	553.39-730.49	514.44-765.50	503.29-735.35	424.38-740.74
MALP-EM	**2008 Scan A**	690.71 (69.12)	716.21 (62.02)	699.34 (69.17)	688.33 (59.38)	673.63 (52.88)
		587.04-871.31	612.20-825.52	621.59-910.91	606.16-874.72	585.90-824.78
	2008 Scan B	689.55 (69.06)	713.93 (61.38)	698.56 (69.70)	686.89 (58.58)	672.44 (54.60)
		589.57-868.85	606.27-821.37	619.76-904.05	609.34-873.16	579.62-824.89
	2011	687.95 (65.36)	706.87 (59.69)	689.91 (70.09)	671.31 (58.60)	658.50 (59.41)
		582.47-849.07	608.90-810.98	609.88-893.67	592.95-854.57	516.56-821.83
FAST	**2008 Scan A**	585.23 (55.70)	594.94 (43.79)	577.22 (50.07)	556.48 (55.90)	544.08 (42.93)
		502.10-723.89	540.27-695.97	500.41-722.76	495.60-737.27	478.45-684.04
	2008 Scan B	581.02 (55.94)	592.16 (42.67)	575.24 (50.79)	553.50 (56.12)	542.90 (44.63)
		495.50-724.37	538.38-682.38	496.78-725.17	490.22-739.64	482.26-685.93
	2011	582.28 (59.24)	583.09 (56.16)	566.74 (54.15)	543.78 (54.71)	510.72 (49.95)
		501.16-742.31	493.89-720.33	461.80-701.45	436.44-672.62	436.44-672.62
FreeSurfer	**2008 Scan A**	600.05 (56.70)	611.24 (53.11)	595.52 (57.52)	576.89 (52.61)	550.74 (46.71)
		527.34-748.92	533.30-728.35	516.49-752.72	505.51-740.96	461.37-686.31
	2008 Scan B	596.88 (56.96)	607.92 (49.83)	592.10 (56.76)	574.23 (52.68)	549.92 (47.50)
		522.55-754.38	537.04-710.30	519.86-749.38	502.84-742.53	465.66-686.73
	2011	597.41 (56.33)	600.51 (53.06)	583.12 (56.77)	559.12 (50.66)	531.77 (47.57)
		526.80-749.19	529.69-715.26	501.32-732.18	487.13-719.53	435.12-676.07
		Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**
			Cortical GM			
SPM8 Unified	**2008 Scan A**	512.77 (68.98)	533.18 (51.24)	516.45 (60.60)	488.49 (55.33)	473.58 (41.97)
		373.73-628.56	478.05-654.50	424.98-638.92	403.99-660.65	422.11-620.20
	2008 Scan B	510.38 (67.91)	533.88 (55.06)	517.41 (58.61)	487.34 (57.17)	472.44 (42.15)
		383.03-632.56	475.51-664.79	442.98-643.06	400.61-674.21	416.66-623.85
	2011	505.72 (68.05)	524.44 (51.07)	510.59 (65.15)	460.43 (50.64)	454.58 (43.52)
		372.24-631.09	464.23-662.46	420.43-629.71	369.75-536.01	400.37-609.44
SPM8 New Segment	**2008 Scan A**	536.43 (62.37)	559.21 (51.87)	541.97 (52.80)	539.74 (50.30)	533.95 (40.36)
		452.89-683.11	487.27-650.20	468.18-690.29	473.83-684.56	469.86-645.25
	2008 Scan B	536.32 (63.46)	558.47 (52.10)	542.70 (53.96)	539.28 (50.29)	533.77 (40.47)
		455.09-689.86	485.55-650.99	466.88-694.94	474.23-685.15	469.95-645.02

	2011	535.67 (63.61)	555.14 (52.30)	538.91 (52.49)	534.52 (48.84)	525.03 (47.10)
		459.16-689.79	484.16-647.05	468.62-685.50	471.89-669.97	404.13-640.42
SPM12	**2008 Scan A**	533.51 (65.41)	552.73 (49.76)	526.07 (48.46)	498.74 (65.09)	480.93 (50.88)
		389.73-666.73	473.23-655.70	460.87-652.17	428.26-705.08	407.50-647.96
	2008 Scan B	533.76 (65.64)	550.80 (47.79)	525.06 (46.78)	497.70 (64.62)	478.57 (51.57)
		394.15-668.75	471.59-645.80	459.22-655.18	423.09-701.96	406.43-648.08
	2011	528.20 (63.27)	542.72 (48.25)	509.88 (51.62)	473.94 (65.99)	455.94 (52.03)
		403.20-662.02	481.15-642.66	449.76-638.26	396.61-670.31	380.94-619.42
ANTS	**2008 Scan A**	495.87 (54.37)	509.72 (44.12)	489.45 (48.51)	477.14 (56.81)	466.91 (45.54)
		389.28-622.33	455.01-611.42	405.19-620.96	409.16-635.36	389.93-604.35
	2008 Scan B	491.64 (56.78)	503.79 (44.09)	487.12 (48.13)	473.74 (54.49)	462.73 (48.51)
		382.09-628.22	449.15-605.89	404.60-619.20	401.01-634.44	376.62-605.35
	2011	488.38 (58.86)	498.77 (44.12)	479.69 (49.48)	461.18 (55.20)	451.16 (53.24)
		387.97-630.09	445.98-594.11	408.16-605.48	388.52-606.14	336.37-596.95
MALP-EM	**2008 Scan A**	530.51 (59.82)	555.69 (50.24)	539.48 (55.25)	533.19 (50.60)	522.94 (43.59)
		437.13-676.95	487.11-657.42	475.98-702.28	468.57-696.92	444.29-646.02
	2008 Scan B	529.34 (60.07)	553.63 (49.68)	539.16 (56.15)	532.20 (49.99)	521.89 (44.94)
		435.90-675.24	482.71-654.95	476.11-696.51	472.30-695.76	439.33-646.89
	2011	527.67 (56.78)	547.99 (48.50)	531.68 (56.27)	519.58 (50.34)	511.12 (47.08)
		443.05-660.08	484.60-646.25	471.96-685.57	444.61-678.12	403.89-641.53
FAST	**2008 Scan A**	469.53 (53.49)	479.12 (35.49)	463.95 (41.39)	447.26 (50.14)	435.76 (36.32)
		374.92-598.60	434.02-561.92	411.23-577.65	390.53-613.65	374.94-553.17
	2008 Scan B	465.54 (54.05)	476.73 (34.58)	462.65 (42.40)	444.91 (50.00)	435.25 (37.43)
		368.99-600.62	428.84-555.23	406.44-579.29	391.75-615.45	378.74-555.11
	2011	466.03 (56.25)	460.96 (38.77)	454.14 (46.04)	427.99 (51.13)	408.29 (43.91)
		378.26-613.36	404.72-544.14	357.59-553.83	353.00-584.79	340.86-539.01
FreeSurfer	**2008 Scan A**	443.98 (46.48)	456.79 (41.53)	438.76 (46.32)	425.94 (48.04)	409.85 (40.17)
		390.66-557.74	401.60-528.21	383.69-569.80	374.81-581.65	331.48-528.49
	2008 Scan B	440.27 (47.23)	453.61 (39.45)	435.40 (46.04)	423.04 (47.57)	412.64 (39.10)
		384.36-563.17	400.90-533.17	374.25-566.91	368.04-583.23	339.36-530.62
	2011	441.53 (46.83)	448.86 (42.55)	428.95 (45.78)	410.99 (46.52)	394.04 (40.47)
		386.74-558.92	390.67-522.61	369.63-548.28	362.78-563.30	318.51-518.13
		Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**
				Frontal GM		
SPM8 Unified	**2008 Scan A**	180.98 (28.52)	189.90 (19.38)	184.76 (23.46)	174.68 (20.13)	170.53 (14.84)
		106.22-230.90	166.41-232.57	151.79-234.20	143.83-241.28	153.66-219.99
	2008 Scan B	180.14 (27.91)	190.17 (21.37)	185.67 (23.05)	174.00 (21.14)	170.15 (15.15)
		109.01-227.44	167.69-238.79	156.12-234.68	143.06-246.39	151.22-221.86
	2011	177.48 (28.29)	186.98 (18.77)	182.80 (24.52)	164.39 (17.52)	163.63 (15.01)
		107.34-228.77	159.57-233.12	148.94-227.21	133.36-189.41	145.85-216.00
SPM8 New Segment	**2008 Scan A**	189.39 (25.88)	199.62 (20.12)	193.43 (20.41)	193.31 (19.55)	191.08 (14.61)
		137.26-245.25	173.83-231.57	164.85-249.67	169.09-252.14	162.67-230.14
	2008 Scan B	189.11 (26.30)	199.37 (20.41)	194.08 (20.83)	193.07 (19.69)	191.11 (14.69)
		136.85-247.62	173.04-232.53	165.29-251.63	169.42-253.10	162.34-229.87
	2011	188.58 (26.26)	197.99 (20.37)	192.02 (20.24)	191.83 (18.95)	187.54 (16.33)
		139.07-246.89	170.54-229.86	163.16-245.60	166.05-246.06	150.15-228.48
SPM12	**2008 Scan A**	187.68 (27.48)	196.61 (18.85)	187.52 (18.47)	178.26 (23.41)	172.80 (17.64)
		110.47-235.58	164.94-232.57	163.93-236.62	150.27-257.28	149.93-229.21
	2008 Scan B	187.69 (27.58)	195.78 (18.45)	187.67 (18.10)	177.69 (23.27)	171.98 (18.08)
		112.51-237.05	165.42-229.56	164.62-238.58	148.61-256.47	148.73-229.93
	2011	185.36 (26.89)	193.05 (17.87)	181.59 (19.53)	169.97 (23.42)	163.97 (17.56)
		115.06-233.49	167.97-226.18	157.27-229.42	135.86-244.01	140.67-219.26
ANTS	**2008 Scan A**	175.75 (22.96)	182.93 (16.80)	175.84 (18.75)	171.78 (21.20)	169.39 (16.87)
		113.82-221.10	164.09-216.83	144.59-228.82	142.45-232.90	134.05-216.09
	2008 Scan B	174.07 (23.98)	180.40 (17.34)	175.53 (18.61)	170.40 (20.55)	167.87 (17.88)
		113.52-225.03	158.33-215.03	145.47-228.63	143.43-232.99	130.33-216.77
	2011	172.02 (25.88)	178.42 (16.72)	172.64 (18.96)	166.58 (20.25)	163.66 (18.72)
		114.78-226.42	155.15-210.80	145.31-220.67	133.86-221.28	127.13-212.46
MALP-EM	**2008 Scan A**	191.50 (26.06)	203.47 (19.52)	197.69 (21.66)	195.12 (19.63)	191.73 (15.84)
		127.79-246.67	178.88-239.10	173.33-259.88	167.21-259.26	167.38-233.41

	2008 Scan B	190.85 (26.08)	202.74 (19.66)	198.10 (21.77)	194.67 (19.57)	191.29 (16.37)
		128.02-245.99	178.61-238.75	174.69-257.95	167.57-259.11	164.06-234.06
	2011	189.58 (24.80)	199.72 (18.69)	194.38 (21.24)	190.07 (19.33)	186.87 (16.63)
		131.21-237.91	175.11-233.90	169.81-247.90	160.37-250.80	154.65-231.24
FAST	**2008 Scan A**	163.82 (22.16)	167.73 (13.40)	163.68 (15.99)	157.54 (19.32)	155.82 (13.72)
		106.03-208.91	148.21-199.57	145.72-210.57	130.85-222.20	138.43-194.28
	2008 Scan B	162.41 (22.53)	166.96 (13.35)	163.76 (16.52)	156.47 (19.39)	155.96 (14.38)
		104.96-211.24	146.69-195.40	144.24-211.70	131.16-223.71	135.82-195.55
	2011	162.78 (24.11)	161.48 (13.90)	160.58 (17.43)	152.57 (17.96)	144.40 (18.81)
		106.91-218.66	138.58-192.57	127.55-199.38	122.21-210.93	102.41-188.63
FreeSurfer	**2008 Scan A**	164.48 (21.29)	171.86 (16.65)	165.86 (19.12)	162.12 (17.95)	156.63 (15.08)
		111.40-206.85	147.76-203.70	140.52-218.79	140.99-221.62	130.10-199.35
	2008 Scan B	163.60 (21.65)	168.27 (17.08)	162.38 (18.21)	156.95 (16.93)	151.26 (14.77)
		111.06-209.39	144.17-202.28	140.12-208.54	134.95-212.67	127.40-195.32
		Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**

Temporal GM

SPM8 Unified	**2008 Scan A**	120.97 (17.93)	127.51 (10.36)	123.21 (14.15)	116.77 (13.11)	115.02 (10.21)
		71.48-146.20	112.63-146.96	99.59-153.25	97.60-158.18	104.52-148.72
	2008 Scan B	120.82 (16.18)	127.08 (10.86)	123.11 (13.13)	116.76 (13.41)	114.77 (10.48)
		80.55-146.52	112.65-147.87	105.97-153.93	96.78-160.93	104.32-149.66
	2011	118.98 (18.33)	125.47 (10.19)	121.96 (16.18)	110.31 (11.60)	110.68 (10.90)
		68.84-146.83	112.77-148.60	100.22-154.86	87.95-124.77	97.95-147.70
SPM8 New Segment	**2008 Scan A**	129.44 (16.30)	136.18 (11.80)	132.04 (13.21)	131.30 (12.11)	131.83 (10.63)
		90.86-161.29	116.62-153.60	114.01-167.99	112.70-165.40	111.46-155.37
	2008 Scan B	129.71 (15.72)	135.91 (11.63)	132.04 (13.02)	131.30 (12.12)	131.78 (10.77)
		97.54-163.58	117.24-153.54	114.95-167.95	112.75-164.55	111.26-155.84
	2011	128.59 (17.20)	135.57 (11.74)	131.54 (13.47)	130.35 (11.77)	129.96 (12.45)
		87.26-163.60	116.93-153.61	116.55-168.75	112.59-163.60	99.14-156.36
SPM12	**2008 Scan A**	127.74 (17.34)	133.53 (10.37)	127.45 (12.42)	121.01 (14.21)	118.64 (11.60)
		78.29-157.50	119.30-154.39	108.11-158.60	103.23-168.38	102.63-155.45
	2008 Scan B	128.16 (16.16)	132.89 (9.85)	127.18 (11.81)	121.03 (14.10)	118.21 (11.86)
		86.97-157.33	120.00-153.77	107.87-158.84	103.65-167.46	102.05-155.69
	2011	125.99 (17.66)	131.45 (9.81)	124.35 (13.39)	115.89 (14.91)	112.99 (12.15)
		76.24-157.23	119.97-152.03	108.65-158.07	99.06-163.07	91.59-150.99
ANTS	**2008 Scan A**	118.96 (15.20)	124.00 (9.77)	118.83 (12.36)	116.12 (12.85)	114.89 (10.87)
		75.80-148.50	107.60-143.07	99.04-150.52	98.43-154.42	93.56-145.76
	2008 Scan B	118.34 (14.41)	122.45 (9.25)	118.22 (11.91)	115.58 (12.34)	113.98 (11.74)
		81.52-148.51	107.41-142.14	99.27-149.84	98.76-153.45	90.55-146.42
	2011	116.05 (16.90)	121.32 (9.31)	116.90 (12.76)	112.58 (12.72)	111.12 (13.20)
		72.29-149.85	107.30-139.22	99.72-149.76	95.98-149.51	81.55-146.39
MALP-EM	**2008 Scan A**	120.61 (15.37)	127.41 (10.43)	123.95 (13.56)	121.95 (12.17)	121.97 (11.22)
		78.39-150.22	109.65-146.95	108.43-162.17	104.51-160.46	102.86-149.50
	2008 Scan B	120.45 (14.50)	126.60 (10.01)	123.70 (13.46)	121.79 (11.88)	121.54 (11.87)
		84.61-150.00	108.76-145.78	108.43-160.55	106.99-160.14	100.46-150.04
	2011	119.15 (15.51)	125.74 (10.25)	122.45 (14.43)	119.08 (12.57)	119.39 (12.37)
		75.05-147.63	108.10-144.92	106.35-161.88	96.72-158.21	90.74-149.86
FAST	**2008 Scan A**	113.51 (14.88)	117.72 (7.96)	113.73 (10.31)	110.45 (11.86)	108.67 (9.25)
		72.60-144.30	104.82-132.29	101.20-140.26	95.34-150.09	94.83-135.59
	2008 Scan B	112.68 (13.70)	116.75 (7.35)	113.20 (10.23)	110.33 (11.62)	108.44 (9.69)
		78.93-143.45	105.28-131.02	100.74-139.91	95.07-149.58	92.85-136.24
	2011	111.88 (15.48)	113.33 (9.02)	112.00 (11.32)	106.09 (12.87)	103.72 (8.91)
		70.35-147.91	98.09-130.50	94.81-137.92	81.21-144.88	92.69-134.05
FreeSurfer	**2008 Scan A**	106.88 (12.25)	111.19 (8.50)	106.40 (11.35)	102.95 (11.84)	100.99 (9.07)
		73.24-129.97	96.37-124.84	92.06-135.95	89.00-142.54	83.29-127.77
	2008 Scan B	105.75 (13.22)	109.48 (8.05)	104.75 (11.77)	99.55 (11.57)	96.92 (9.22)
		67.72-129.51	97.48-123.75	91.29-135.46	86.87-139.14	77.98-125.87
		Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**

Parietal GM

SPM8 Unified	**2008 Scan A**	115.01 (13.95)	117.89 (15.58)	121.34 (12.66)	117.97 (14.25)	110.74 (12.89)
		89.95-152.64	95.11-146.54	105.21-152.64	97.86-147.60	89.95-144.75
	2008 Scan B	116.94 (15.88)	121.60 (13.49)	117.88 (13.77)	110.43 (13.22)	106.98 (9.75)
		95.27-147.28	105.22-154.63	100.64-148.25	89.12-147.97	91.54-138.53
	2011	116.75 (15.11)	119.05 (12.57)	116.13 (14.61)	104.72 (12.72)	103.14 (10.00)
		95.26-148.23	102.71-153.81	96.62-144.24	83.10-129.14	88.11-134.68
SPM8 New Segment	**2008 Scan A**	124.02 (12.62)	123.51 (15.65)	128.21 (12.34)	124.11 (12.47)	122.90 (11.91)
		101.06-160.24	101.20-160.24	108.70-147.85	108.55-158.98	105.71-151.24
	2008 Scan B	123.41 (15.83)	128.13 (12.46)	124.28 (13.19)	122.76 (11.89)	121.25 (10.21)
		102.32-161.06	108.37-148.48	108.28-159.96	105.66-151.80	102.32-146.77
	2011	123.11 (12.67)	123.96 (15.75)	127.17 (12.40)	123.34 (12.11)	121.58 (11.56)
		90.36-162.22	103.50-162.22	107.32-148.16	108.48-157.08	103.84-148.09
SPM12	**2008 Scan A**	116.99 (14.45)	121.34 (14.76)	124.82 (12.22)	118.79 (11.36)	112.15 (15.47)
		86.12-153.77	101.06-153.41	103.00-149.48	104.84-148.29	94.80-153.77
	2008 Scan B	121.14 (15.15)	124.63 (11.66)	118.42 (11.09)	111.85 (15.41)	107.35 (12.16)
		100.60-153.20	103.29-148.02	104.39-148.85	94.57-153.48	86.20-143.52
	2011	113.40 (15.10)	120.62 (14.26)	122.57 (11.57)	114.89 (11.71)	106.40 (15.68)
		84.14-153.40	99.82-153.40	105.86-147.62	101.51-143.97	89.46-145.79
ANTS	**2008 Scan A**	111.28 (12.40)	114.25 (13.21)	116.18 (11.09)	111.81 (11.30)	108.56 (13.48)
		87.01-146.34	96.43-146.34	100.40-140.44	94.34-143.47	90.78-139.49
	2008 Scan B	112.90 (13.78)	115.21 (10.80)	111.23 (11.70)	107.66 (12.88)	104.58 (11.42)
		89.78-146.87	99.01-139.28	94.38-142.84	90.95-139.47	86.25-135.21
	2011	108.72 (12.86)	113.70 (13.25)	113.84 (10.84)	109.06 (11.38)	104.74 (13.43)
		75.84-148.00	94.40-148.00	99.54-137.42	95.15-138.12	86.88-132.96
MALP-EM	**2008 Scan A**	128.68 (13.28)	128.51 (15.76)	133.19 (12.95)	129.25 (14.00)	127.62 (11.97)
		100.96-171.44	107.34-166.99	111.67-157.42	114.14-171.44	112.08-163.43
	2008 Scan B	128.36 (15.82)	132.99 (12.88)	129.16 (14.58)	127.40 (11.85)	124.68 (11.23)
		108.56-165.23	110.71-157.08	113.72-169.92	112.79-163.15	101.35-153.79
	2011	126.98 (13.19)	128.34 (15.14)	131.86 (12.51)	127.44 (13.74)	124.78 (11.97)
		94.89-166.33	105.49-163.52	110.97-155.82	112.28-166.33	108.64-158.58
FAST	**2008 Scan A**	105.40 (11.45)	108.77 (13.33)	110.26 (9.35)	106.65 (9.91)	102.51 (11.85)
		79.20-142.62	92.09-142.62	97.87-131.35	93.53-133.99	86.09-135.15
	2008 Scan B	107.71 (13.41)	109.93 (9.03)	106.28 (10.41)	101.73 (11.82)	98.65 (9.35)
		90.41-142.31	97.17-128.13	92.49-133.95	85.67-135.66	80.43-123.57
	2011	102.59 (12.27)	108.67 (14.04)	106.62 (9.55)	103.91 (10.16)	98.60 (12.39)
		77.32-147.36	90.60-147.36	91.76-125.41	83.14-126.64	81.72-128.59
FreeSurfer	**2008 Scan A**	105.34 (12.16)	106.83 (10.30)	104.04 (11.30)	99.33 (11.99)	95.51 (10.42)
		88.40-133.55	91.71-125.28	90.43-138.07	85.11-133.53	73.41-122.91
	2008 Scan B	104.84 (12.06)	105.29(10.37)	100.97 (11.16)	95.56 (11.65)	91.82 (10.41)
		88.65-134.08	88.88-124.18	85.43-130.83	82.21-129.22	72.02-119.97
		Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**

Occipital GM

SPM8 Unified	**2008 Scan A**	81.47 (9.82)	82.01 (10.19)	78.28 (8.65)	74.65 (10.20)	69.51 (9.31)
		62.78-101.30	69.74-107.25	64.63-94.07	58.66-100.64	56.92-97.25
	2008 Scan B	80.90 (10.05)	82.56 (10.57)	78.43 (8.51)	74.53 (10.38)	69.25 (9.30)
		63.16-102.25	70.01-108.06	64.57-93.39	59.24-102.89	56.77-97.68
	2011	80.94 (9.33)	80.63 (10.91)	77.41 (9.57)	69.83 (10.00)	66.12 (9.92)
		63.60-101.97	68.29-111.57	63.09-92.09	52.28-90.05	53.65-95.33
SPM8 New Segment	**2008 Scan A**	83.36 (8.97)	83.61 (9.48)	80.94 (7.14)	81.02 (8.10)	78.34 (7.83)
		71.69-103.91	69.29-104.35	67.62-100.31	70.00-101.82	65.28-97.80
	2008 Scan B	83.24 (8.91)	83.44 (9.48)	80.80 (7.30)	81.03 (7.96)	78.19 (7.79)
		70.93-105.39	69.11-103.94	66.53-101.79	70.38-102.00	65.19-97.74
	2011	83.90 (9.05)	82.83 (9.79)	80.60 (6.94)	79.59 (8.07)	76.59 (9.57)
		72.03-105.30	69.23-102.81	68.49-99.86	69.17-99.05	55.48-97.59
SPM12	**2008 Scan A**	84.39 (8.73)	84.49 (10.05)	79.53 (6.62)	75.13 (12.29)	69.92 (10.88)
		71.15-106.01	71.30-108.76	68.82-94.83	57.08-108.49	55.83-102.09
	2008 Scan B	84.26 (8.82)	84.27 (9.65)	79.04 (6.40)	75.01 (12.18)	69.33 (10.96)
		71.39-106.62	70.86-106.87	68.60-95.10	57.19-107.75	55.98-101.75
	2011	83.82 (8.23)	82.55 (10.57)	76.51 (7.48)	69.92 (12.49)	65.13 (11.60)
		68.72-103.74	66.99-107.84	65.74-92.72	52.90-101.33	50.50-96.45

ANTS	**2008 Scan A**	76.52 (6.94)	75.39 (9.09)	71.99 (6.79)	69.77 (10.13)	66.02 (9.93)
		66.52-94.94	64.84-99.30	56.89-86.68	57.48-94.04	53.06-92.53
	2008 Scan B	75.83 (7.04)	74.56 (9.04)	71.18 (6.53)	69.36 (9.53)	65.34 (10.12)
		64.54-95.67	63.80-97.44	56.08-85.74	57.71-94.14	52.37-92.35
	2011	76.21 (6.87)	73.96 (9.38)	69.97 (6.90)	66.51 (9.72)	63.19 (11.42)
		65.75-94.43	62.28-94.63	56.71-84.46	53.45-88.89	42.58-91.34
MALP-EM	**2008 Scan A**	80.33 (8.04)	80.95 (9.45)	78.11 (7.29)	77.90 (8.10)	74.22 (8.52)
		70.21-100.69	67.11-102.78	66.93-100.00	65.76-100.17	60.68-95.42
	2008 Scan B	80.05 (7.80)	80.67 (9.33)	77.68 (7.38)	77.79 (7.83)	74.15 (8.49)
		69.42-101.04	66.51-102.08	65.48-98.47	65.93-100.01	61.67-95.56
	2011	80.88 (7.80)	80.03 (9.18)	76.96 (7.79)	75.12 (8.06)	72.12 (9.52)
		70.39-98.78	67.27-100.26	65.25-98.43	64.76-97.58	54.74-94.83
FAST	**2008 Scan A**	73.46 (7.46)	72.38 (8.19)	69.23 (6.11)	66.32 (9.25)	62.13 (8.40)
		59.83-92.11	63.11-91.62	60.51-81.93	53.29-91.62	49.91-85.26
	2008 Scan B	72.75 (7.29)	72.12 (7.95)	68.73 (6.09)	65.94 (9.16)	61.85 (8.42)
		57.90-92.49	62.27-90.17	59.24-82.21	53.22-91.90	50.62-85.35
	2011	72.92 (7.86)	68.72 (9.47)	66.89 (8.31)	60.66 (11.95)	55.42 (12.33)
		60.31-91.98	49.13-87.01	41.90-79.72	31.04-86.52	34.97-82.97
FreeSurfer	**2008 Scan A**	66.30 (7.74)	66.07 (7.47)	61.89 (6.19)	59.68 (8.89)	54.85 (8.43)
		57.48-85.09	56.14-83.26	52.82-76.74	45.74-83.49	42.19-76.40
	2008 Scan B	66.26 (7.12)	65.00 (7.87)	60.02 (6.45)	57.10 (8.36)	52.23 (8.95)
		56.46-83.60	54.31-81.47	49.60-72.51	45.10-80.61	40.04-74.45
		Controls (N=20)	**PreHD-A (N=20)**	**PreHD-B (N=20)**	**HD1 (N=20)**	**HD2 (N=20)**

Insula GM

SPM8 Unified	**2008 Scan A**	81.47 (9.82)	82.01 (10.19)	78.28 (8.65)	74.65 (10.20)	69.51 (9.31)
		62.78-101.30	69.74-107.25	64.63-94.07	58.66-100.64	56.92-97.25
	2008 Scan B	80.90 (10.05)	82.56 (10.57)	78.43 (8.51)	74.53 (10.38)	69.25 (9.30)
		63.16-102.25	70.01-108.06	64.57-93.39	59.24-102.89	56.77-97.68
	2011	80.94 (9.33)	80.63 (10.91)	77.41 (9.57)	69.83 (10.00)	66.12 (9.92)
		63.60-101.97	68.29-111.57	63.09-92.09	52.28-90.05	53.65-95.33
SPM8 New Segment	**2008 Scan A**	83.36 (8.97)	83.61 (9.48)	80.94 (7.14)	81.02 (8.10)	78.34 (7.83)
		71.69-103.91	69.29-104.35	67.62-100.31	70.00-101.82	65.28-97.80
	2008 Scan B	83.24 (8.91)	83.44 (9.48)	80.80 (7.30)	81.03 (7.96)	78.19 (7.79)
		70.93-105.39	69.11-103.94	66.53-101.79	70.38-102.00	65.19-97.74
	2011	83.90 (9.05)	82.83 (9.79)	80.60 (6.94)	79.59 (8.07)	76.59 (9.57)
		72.03-105.30	69.23-102.81	68.49-99.86	69.17-99.05	55.48-97.59
SPM12	**2008 Scan A**	84.39 (8.73)	84.49 (10.05)	79.53 (6.62)	75.13 (12.29)	69.92 (10.88)
		71.15-106.01	71.30-108.76	68.82-94.83	57.08-108.49	55.83-102.09
	2008 Scan B	84.26 (8.82)	84.27 (9.65)	79.04 (6.40)	75.01 (12.18)	69.33 (10.96)
		71.39-106.62	70.86-106.87	68.60-95.10	57.19-107.75	55.98-101.75
	2011	83.82 (8.23)	82.55 (10.57)	76.51 (7.48)	69.92 (12.49)	65.13 (11.60)
		68.72-103.74	66.99-107.84	65.74-92.72	52.90-101.33	50.50-96.45
ANTS	**2008 Scan A**	76.52 (6.94)	75.39 (9.09)	71.99 (6.79)	69.77 (10.13)	66.02 (9.93)
		66.52-94.94	64.84-99.30	56.89-86.68	57.48-94.04	53.06-92.53
	2008 Scan B	75.83 (7.04)	74.56 (9.04)	71.18 (6.53)	69.36 (9.53)	65.34 (10.12)
		64.54-95.67	63.80-97.44	56.08-85.74	57.71-94.14	52.37-92.35
	2011	76.21 (6.87)	73.96 (9.38)	69.97 (6.90)	66.51 (9.72)	63.19 (11.42)
		65.75-94.43	62.28-94.63	56.71-84.46	53.45-88.89	42.58-91.34
MALP-EM	**2008 Scan A**	80.33 (8.04)	80.95 (9.45)	78.11 (7.29)	77.90 (8.10)	74.22 (8.52)
		70.21-100.69	67.11-102.78	66.93-100.00	65.76-100.17	60.68-95.42
	2008 Scan B	80.05 (7.80)	80.67 (9.33)	77.68 (7.38)	77.79 (7.83)	74.15 (8.49)
		69.42-101.04	66.51-102.08	65.48-98.47	65.93-100.01	61.67-95.56
	2011	80.88 (7.80)	80.03 (9.18)	76.96 (7.79)	75.12 (8.06)	72.12 (9.52)
		70.39-98.78	67.27-100.26	65.25-98.43	64.76-97.58	54.74-94.83
FAST	**2008 Scan A**	73.46 (7.46)	72.38 (8.19)	69.23 (6.11)	66.32 (9.25)	62.13 (8.40)
		59.83-92.11	63.11-91.62	60.51-81.93	53.29-91.62	49.91-85.26
	2008 Scan B	72.75 (7.29)	72.12 (7.95)	68.73 (6.09)	65.94 (9.16)	61.85 (8.42)
		57.90-92.49	62.27-90.17	59.24-82.21	53.22-91.90	50.62-85.35
	2011	72.92 (7.86)	68.72 (9.47)	66.89 (8.31)	60.66 (11.95)	55.42 (12.33)
		60.31-91.98	49.13-87.01	41.90-79.72	31.04-86.52	34.97-82.97

FreeSurfer	**2008 Scan A**	66.30 (7.74)	66.07 (7.47)	61.89 (6.19)	59.68 (8.89)	54.85 (8.43)
		57.48-85.09	56.14-83.26	52.82-76.74	45.74-83.49	42.19-76.40
	2008 Scan B	66.26 (7.12)	65.00 (7.87)	60.02 (6.45)	57.10 (8.36)	52.23 (8.95)
		56.46-83.60	54.31-81.47	49.60-72.51	45.10-80.61	40.04-74.45

3.3.3.2. Reliability measures

3.3.3.2.1. Controls

For controls, ICC values for baseline scan pairs were above 0.90 for total, cortical and lobular GM volume using each segmentation tool (Table 3.7) Mean repeatability (indexing variability), was lowest in total GM for all techniques, ranging from 0.35% (SPM8 New Segment) to 1.36% (FreeSurfer), as shown in Table 3.7. CGM showed only slightly higher variability than total GM for all techniques. Lobular regions generally had higher repeatability values than total and cortical GM, indicating more variability between the lobular volumes for the first and second back-to-back baseline scans.

Spearman's Rho correlations showed that there were strong relationships between the volumes extracted using the seven different tools, with most values above 0.90 (Table 3.8). Correlations were higher for CGM than for total GM. Spearman's Rho for lobular regions are not shown, but relationships > .75 were seen for all measures for controls across all regions, with most results > .90, indicating a high level of agreement between tools. Overall, SPM8 Unified Segment showed the lowest relationships with other measures.

Table 3.7. Intraclass correlation coefficients and confidence intervals for control participants for back-to-back segmentations of total GM, cortical GM, frontal lobe GM, temporal lobe GM, parietal lobe GM, occipital lobe GM, and insula GM included in the current study. Repeatability values are also displayed for back-to-back segmentations of total GM, cortical GM, frontal lobe GM, temporal lobe GM, parietal lobe GM, occipital lobe GM, and insula GM for all control participants included in the current study, showing means, standard deviations and ranges.

	Intraclass Correlations Confidence Intervals						
	SPM8 Unified Segment	**SPM8 New Segment**	**SPM12 Segment**	**ANTs Atropos**	**MALP-EM**	**FSL FAST**	**FreeSurfer**
Total GM	.994	.999	.997	.982	.998	.986	.978
	.985-.998	.997-1.000	.993-.999	.951-.993	.995-.999	.960-.995	.947-.991
Cortical GM	.994	.999	.997	.985	.998	.988	.967
	.985-.998	.998-1.000	.993-.999	.958-.994	.995-.999	.964-.996	.918-.987
Frontal Lobe	0.996	0.999	0.997	0.983	0.997	0.989	0.960
	0.990-0.998	0.997-.000	0.991-0.999	0.955-0.993	0.993-0.999	0.969-0.996	0.902-0.984
Temporal	0.989	0.994	0.992	0.990	0.994	0.985	.975
Lobe	0.973-0.996	0.986-0.998	0.980-0.997	0.975-0.996	0.984-0.997	0.963-0.994	0.936-0.990
Parietal Lobe	0.995	0.997	0.996	0.976	0.996	0.984	0.956
	0.986-0.998	0.993-0.999	0.990-0.998	0.931-0.991	0.990-0.998	0.954-0.994	0.886-0.983
Occipital	0.994	0.993	0.995	0.971	0.992	0.978	0.962
Lobe	0.985-0.998	0.984-0.997	0.988-0.998	0.922-0.989	0.981-0.997	0.936-0.992	0.906-0.985
Insula	0.977	0.979	0.979	0.982	0.985	0.979	0.975
	0.941-0.991	0.948-0.992	0.948-0.992	0.955-0.993	0.962-0.994	0.947-0.991	0.938-0.990
	Mean Repeatability (Standard Deviation) Range						
	SPM8 Unified Segment	**SPM8 New Segment**	**SPM12 Segment**	**ANTs Atropos**	**MALP-EM**	**FSL FAST**	**FreeSurfer**
Total GM	1.08	0.35	0.69	1.14	0.41	0.91	1.36
	(0.82)	(0.34)	(0.48)	(1.65)	(0.53)	(1.33)	(1.43)
	0.18-2.88	0.01-1.00	0.12-1.77	0.00-6.44	0.02-2.07	0.05-4.99	0.05-5.77
Cortical GM	1.06	0.36	0.79	1.23	0.49	1.14	1.82
	(0.82)	(0.25)	(0.50)	(1.71)	(0.57)	(1.36)	(2.09)
	0.02-2.72	0.00-0.98	0.07-1.96	0.03-6.81	0.01-2.18	0.03-4.81	0.06-8.05
Frontal Lobe	1.13	0.50	1.09	1.50	0.67	1.43	2.33
	(0.93)	(0.43)	(0.67)	(2.28)	(0.74)	(1.49)	(3.05)
	0.00-3.09	0.05-1.48	0.20-2.45	0.02-9.65	0.00-2.67	0.01-5.72	0.01-11.04
Temporal	1.50	0.82	1.06	1.25	0.90	1.54	1.80
Lobe	(2.63)	(1.52)	(2.27)	(1.86)	(1.71)	(2.07)	(2.36)
	0.03-11.93	0.02-7.09	0.04-10.50	0.00-7.27	0.00-7.62	0.02-8.36	0.22-9.56
Parietal Lobe	1.07	0.59	0.88	1.57	0.80	1.29	2.21
	(0.88)	(0.79)	(0.87)	(2.35)	(0.75)	(1.91)	(2.79)
	0.01-3.29	0.04-3.72	0.14-3.78	0.01-7.25	0.00-2.74	0.11-6.03	0.10-8.84
Occipital	1.02	0.85	0.88	1.36	0.75	1.29	2.23
Lobe	(0.96)	(0.82)	(0.63)	(1.79)	(0.86)	(1.63)	(1.86)
	0.07-3.53	0.04-3.29	0.15-2.51	0.07-6.57	0.07-3.93	0.03-6.00	0.00-6.56
Insula	1.73	1.48	1.44	1.51	1.31	1.58	2.07
	(2.42)	(1.66) 0.05-	(2.13) 0.22-	(1.76) 0.21-	(1.88)	(2.21)	(1.91)
	0.08-10.97	7.91	9.90	8.32	0.01-8.44	0.09-10.55	0.09-7.43

Table 3.8. Spearman's rank correlation for control participants for total GM and cortical GM.

	SPM8 Unified Segment	SPM8 New Segment	SPM12 Segment	ANTs Atropos	MALP-EM	FSL FAST
Total GM						
SPM8 Unified Segment	1					
SPM8 New Segment	0.761	1				
SPM12 Segment	0.904	0.857	1			
ANTs Atropos	0.788	0.920	0.929	1		
MALP-EM	0.812	0.958	0.928	0.967	1	
FSL FAST	0.867	0.896	0.956	0.929	0.944	1
FreeSurfer	0.874	0.904	0.884	0.874	0.905	0.920
Cortical GM						
SPM8 Unified Segment	1					
SPM8 New Segment	0.798	1				
SPM12 Segment	0.918	0.971	1			
ANTs Atropos	0.809	0.976	0.934	1		
MALP-EM	0.861	0.932	0.953	0.974	1	
FSL FAST	0.897	0.932	0.947	0.944	0.967	1
FreeSurfer	0.844	0.910	0.893	0.919	0.962	0.956

3.3.3.2.2. HD gene-carrier participants

Reliability for total, cortical and lobular GM as measured by ICC was above 0.90 for most tools across all regions and the disease subgroups (Table 3.9 and Appendix 1, Table 1).

Repeatability values were more variable than ICC values. Repeatability was lower with increasing disease stage for total and cortical GM (Table 3.10). For individual lobes, it was more variable across disease stages, lobes, and tools, but showed a small range of mean values (Appendix 1, Table 2). Values ranged from 0.37% (parietal lobe volume measured by MALP-EM in HD1 participants) to 2.72% (insula volume measured by FreeSurfer in PreHD-B participants).

For total and cortical GM Spearman's correlation between measures tended to be lower for the HD2 group, indicating that the techniques perform differently on more atrophied brains (Table 3.11 and Table 3.12). SPM8 Unified Segment again had lower values than other techniques for Spearman's Rho, especially with SPM8 New Segment in HD2 participants, whereby correlations of 0.441 and 0.411 were seen for total and cortical GM respectively. For lobular regions, the relationships between measures were generally lower than those in control participants, with more correlations between 0.7-0.9 (Appendix 1, Tables 3-8).

Table 3.9. (A) Intraclass correlation coefficients and confidence intervals for HD participants for all tools measuring total GM volume in back-to-back 2008 scans (B) Intraclass correlation coefficients and confidence intervals for HD participants for all tools measuring CGM volume for back-to-back 2008 scans.

	PreHD-A (N=20)	PreHD-B (N=20)	HD1 (N=20)	HD2 (N=20)
		Total GM Intraclass Correlations Confidence intervals		
SPM8 Unified	0.993 0.982-0.997	0.977 0.943-0.991	0.984 0.960-0.994	0.989 0.973-0.996
SPM8 New Segment	0.999 0.998-10.000	0.999 0.998-10.000	0.999 0.998-10.000	0.999 0.998-10.000
SPM12	0.989 0.972-0.995	0.990 0.974-0.996	0.997 0.994-0.999	0.994 0.982-0.998
Atropos	0.975 0.912-0.991	0.994 0.981-0.998	0.989 0.970-0.996	0.986 0.938-0.995
MALP-EM	0.996 0.989-0.998	0.998 0.996-0.999	0.999 0.996-0.999	0.996 0.990-0.998
FAST	0.989 0.970-0.996	0.997 0.989-0.999	0.995 0.982-0.998	0.994 0.985-0.998
FreeSurfer	0.988 0.967-0.995	0.988 0.968-0.995	0.992 0.979-0.997	0.993 0.983-0.997
		Cortical GM Intraclass Correlations Confidence intervals		
SPM8 Unified	0.993 0.982-0.997	0.979 0.947-0.992	0.987 0.967-0.995	0.991 0.977-0.996
SPM8 New Segment	0.999 0.998-10.000	0.999 0.998-10.000	0.999 0.998-10.000	0.999 0.998-10.000
SPM12	0.988 0.971-0.995	0.991 0.978-0.996	0.998 0.995-0.999	0.995 0.984-0.998
Atropos	0.970 0.893-0.990	0.995 0.984-0.998	0.991 0.974-0.997	0.990 0.947-0.997
MALP-EM	0.995 0.986-0.998	0.998 0.994-0.999	0.999 0.997-10.000	0.996 0.991-0.999
FAST	0.986 0.961-0.994	0.996 0.989-0.998	0.995 0.985-0.998	0.995 0.987-0.998
FreeSurfer	0.985 0.957-0.994	0.983 0.955-0.993	0.989 0.970-0.996	0.992 0.981-0.997

Table 3.10. (A) Repeatability values for back-to-back segmentations of total GM for all HD participants included in the current study, showing means, standard deviations, and ranges (B) Repeatability values for back-to-back segmentations of CGM for all HD participants included in the current study, showing means, standard deviations, and ranges.

	SPM8 Unified	SPM8 New	SPM12	Atropos	MALP-EM	FAST	FreeSurfer
				A) Total GM			
PreHD-A Total GM	0.89 (0.80) 0.01-3.05	0.27 (0.19) 0.02-0.80	0.97 (0.86) 0.01-2.93	1.26 (1.63) 0.07-5.61	0.58 (0.53) 0.02-2.22	0.82 (0.67) 0.00-2.42	0.90 (0.87) 0.03-2.69
PreHD-B Total GM	1.61 (1.83) 0.24-7.03	0.28 (0.22) 0.02-0.70	0.83 (0.88) 0.02-3.15	0.78 (0.72) 0.07-2.45	0.40 (0.40) 0.01-1.37	0.54 (0.51) 0.02-1.75	1.15 (0.98) 0.03-3.85
HD1 Total GM	1.13 (1.33) 0.04-6.22	0.29 (0.24) 0.01-0.75	0.64 (0.54) 0.08-1.80	1.00 (1.27) 0.02-4.45	0.39 (0.27) 0.08-1.20	0.82 (0.68) 0.13-2.20	0.91 (0.74) 0.15-3.55
HD2 Total GM	0.94 (0.77) 0.07-2.90	0.22 (0.12) 0.03-0.37	0.86 (0.71) 0.03-3.20	1.24 (1.33) 0.02-5.08	0.57 (0.50) 0.01-1.87	0.65 (0.70) 0.00-3.06	0.80 (0.72) 0.05-2.54
				B) Cortical GM			
PreHD-A Cortical GM	0.91 (0.76) 0.05-2.82	0.33 (0.22) 0.04-0.75	1.03 (0.89) 0.10-3.32	1.31 (1.77) 0.09-6.19	0.67 (0.64) 0.05-2.72	1.01 (0.70) 0.02-2.41	1.08 (1.03) 0.03-3.15
PreHD-B Cortical GM	1.61 (1.90) 0.09-7.24	0.31 (0.27) 0.00-1.10	0.86 (0.81) 0.06-2.91	0.77 (0.69) 0.15-2.60	0.47 (0.47) 0.02-1.70	0.65 (0.53) 0.00-1.86	1.49 (1.28) 0.23-5.22
HD1 Cortical GM	1.14 (1.44) 0.11-6.76	0.35 (0.25) 0.03-0.88	0.67 (0.55) 0.06-1.71	1.02 (1.09) 0.00-3.69	0.37 (0.31) 0.03-1.43	0.87 (0.66) 0.11-2.35	1.30 (1.11) 0.27-5.16
HD2 Cortical GM	0.96 (0.78) 0.04-2.74	0.22 (0.16) 0.02-0.61	0.86 (0.70) 0.02-3.10	1.13 (1.11) 0.04-3.47	0.61 (0.49) 0.09-1.85	0.66 (0.63) 0.11-2.77	1.04 (0.96) 0.04-3.17

Table 3.11. Spearman's ranked correlation for segmentations of total GM for all HD participants included in the current study.

	SPM8 Unified Segment	SPM8 New Segment	SPM12 Segment	ANTs Atropos	MALP-EM	FSL FAST
			PreHD-A			
SPM8 Unified Segment	1					
SPM8 New Segment	0.777	1				
SPM12 Segment	0.896	0.666	1			
ANTs Atropos	0.732	0.833	0.737	1		
MALP-EM	0.808	0.950	0.749	0.844	1	
FSL FAST	0.741	0.785	0.770	0.926	0.874	1
FreeSurfer	0.865	0.959	0.752	0.805	0.952	0.802
			PreHD-B			
SPM8 Unified Segment	1					
SPM8 New Segment	0.770	1				
SPM12 Segment	0.932	0.889	1			
ANTs Atropos	0.726	0.934	0.851	1		
MALP-EM	0.842	0.964	0.941	0.919	1	
FSL FAST	0.883	0.848	0.917	0.857	0.878	1
FreeSurfer	0.871	0.925	0.952	0.838	0.938	0.887
			HD1			
SPM8 Unified Segment	1					
SPM8 New Segment	0.737	1				
SPM12 Segment	0.857	0.672	1			
ANTs Atropos	0.847	0.818	0.779	1		
MALP-EM	0.793	0.955	0.702	0.886	1	
FSL FAST	0.845	0.731	0.815	0.917	0.841	1
FreeSurfer	0.884	0.857	0.875	0.829	0.895	0.826
			HD2			
SPM8 Unified Segment	1					
SPM8 New Segment	0.441	1				
SPM12 Segment	0.758	0.555	1			
ANTs Atropos	0.644	0.820	0.671	1		
MALP-EM	0.633	0.917	0.719	0.860	1	
FSL FAST	0.827	0.738	0.768	0.908	0.853	1
FreeSurfer	0.605	0.802	0.796	0.681	0.883	0.726

Table 3.12. Spearman's ranked correlation for segmentations of cortical GM for all HD participants included in the current study.

	SPM8 Unified Segment	SPM8 New Segment	SPM12 Segment	ANTs Atropos	MALP-EM	FSL FAST
PreHD-A						
SPM8 Unified Segment	1					
SPM8 New Segment	0.815	1				
SPM12 Segment	0.883	0.722	1			
ANTs Atropos	0.865	0.901	0.844	1		
MALP-EM	0.836	0.970	0.755	0.886	1	
FSL FAST	0.848	0.848	0.847	0.920	0.874	1
FreeSurfer	0.868	0.955	0.811	0.917	0.973	0.923
PreHD-B						
SPM8 Unified Segment	1					
SPM8 New Segment	0.747	1				
SPM12 Segment	0.926	0.863	1			
ANTs Atropos	0.762	0.943	0.896	1		
MALP-EM	0.830	0.974	0.922	0.947	1	
FSL FAST	0.818	0.934	0.925	0.916	0.955	1
FreeSurfer	0.887	0.836	0.935	0.848	0.874	0.925
HD1						
SPM8 Unified Segment	1					
SPM8 New Segment	0.734	1				
SPM12 Segment	0.868	0.716	1			
ANTs Atropos	0.869	0.898	0.844	1		
MALP-EM	0.735	0.973	0.711	0.914	1	
FSL FAST	0.820	0.764	0.862	0.902	0.782	1
FreeSurfer	0.820	0.893	0.898	0.910	0.901	0.854
HD2						
SPM8 Unified Segment	1					
SPM8 New Segment	0.411	1				
SPM12 Segment	0.826	0.576	1			
ANTs Atropos	0.692	0.814	0.741	1		
MALP-EM	0.638	0.910	0.749	0.869	1	
FSL FAST	0.803	0.699	0.835	0.919	0.860	1
FreeSurfer	0.729	0.729	0.887	0.832	0.868	0.880

3.3.3.3. Longitudinal results

3.3.3.3.1. Total GM volume

Total GM volume change (as a percentage of baseline volume) for all tools was smaller in controls than that of the HD gene-carrier groups (Table 3.13 and Figure 3.15). However when total GM volume change within each HD group was statistically compared to controls, MALP-EM and FreeSurfer were the only two tools that detected significantly greater change in all disease groups. All other tools detected significantly greater change in HD1 and HD2 compared to controls, with SPM12 and FAST also showing greater change in PreHD-B compared to controls.

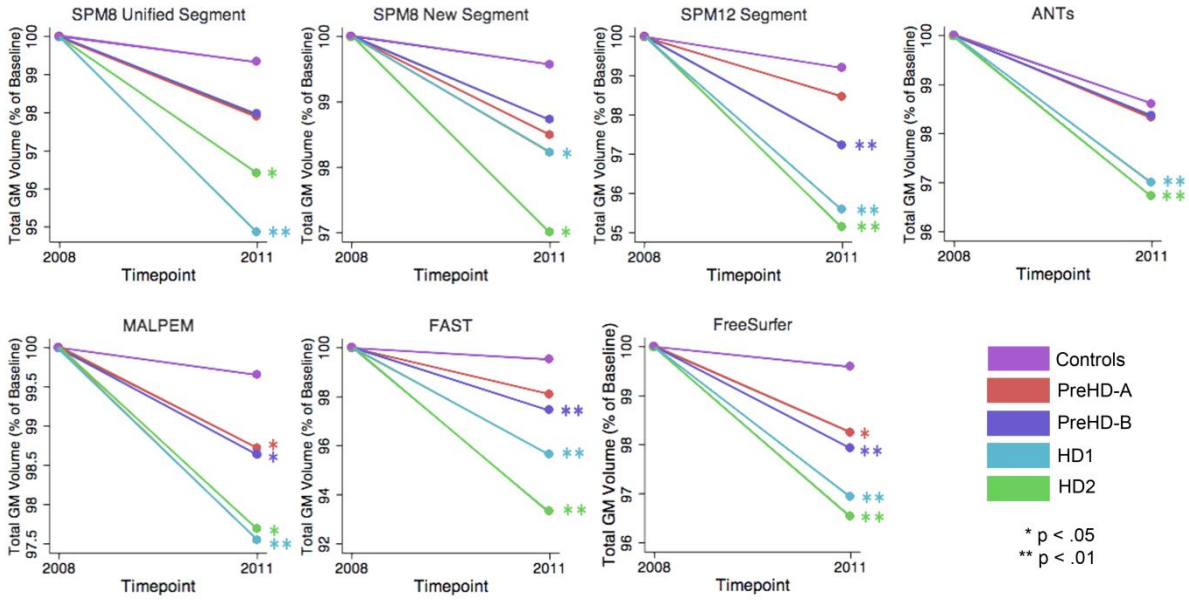


Figure 3.15. Mean values for all tools and groups showing 2011 volume as a percentage of baseline volume in total GM. Significant change difference relative to controls after controlling for age, sex and site are represented by * p<.05, **p<.01.

Table 3.13. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in total GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified Segment	**% Decrease 2008 to 2011**	0.67 (3.05) -5.42-7.93	2.11 (3.06) -2.13-11.07	2.08 (4.53) -7.22-14.38	5.15 (6.97) -7.37-30.87	3.58 (2.06) -0.53-6.28
	Significant difference	-	1.12 (-0.75-3.00) p = 0.239	0.62 (-0.55-1.79) p = 0.299	1.71 (0.65-2.76) p = 0.002	0.53 (0.06-1.00) p = 0.027
SPM8 New Segment	**% Decrease 2008 to 2011**	0.43 (2.41) -7.93-3.23	1.51 (1.09) -0.32-3.30	1.28 (0.76) -0.26-2.60	1.78 (1.04) 0.14-4.12	3.00 (3.39) -0.28-16.62
	Significant difference	-	1.05 (-0.05-2.16) p = 0.062	0.44 (-0.07-0.95) p = 0.089	0.41 (0.06-0.76) p = 0.021	0.59 (0.09-1.08) p = 0.021
SPM12 Segment	**% Decrease 2008 to 2011**	0.80 (2.29) -4.79-3.97	1.54 (3.61) -7.49-7.29	2.77 (1.64) 0.28-5.89	4.42 (2.06) 0.82-7.80	4.87 (2.50) 0.98-9.65
	Significant difference	-	0.43 (-1.40-2.26) p = 0.646	1.00 (0.42-1.58) p = 0.001	1.21 (0.78-1.65) p = 0.000	0.78 (0.36-1.19) p = 0.000
ANTs	**% Decrease 2008 to 2011**	1.39 (2.38) -0.83-9.69	1.67 (2.57) -2.79-6.97	1.65 (2.18) -2.17-7.17	3.00 (1.96) 0.10-7.09	3.28 (4.26) -1.40-18.87
	Significant difference	-	0.59 (-0.90-2.08) p = 0.435	0.20 (-0.44-0.84) p = 0.544	0.58 (0.16-1.00) p = 0.006	0.75 (0.18-1.31) p = 0.009
MALP-EM	**% Decrease 2008 to 2011**	0.35 (1.36) -3.81-2.55	1.28 (1.29) -1.28-4.68	1.37 (1.20) -1.08-3.59	2.45 (2.36) -3.34-9.15	2.31 (2.72) -1.76-11.83
	Significant difference	-	0.89 (0.06-1.72) p = 0.036	0.47 (0.10-0.85) p = 0.013	0.65 (0.28-1.02) p = 0.001	0.39 (0.02-0.75) p = 0.039
FAST	**% Decrease 2008 to 2011**	0.49 (3.57) -7.91-8.88	1.90 (7.83) -26.04-11.48	2.57 (3.02) -0.72-12.70	4.38 (4.47) -4.55-12.72	6.70 (5.81) 0.32-18.59
	Significant difference	-	0.30 (-2.62-3.23) p = 0.839	1.15 (0.44-1.86) p = 0.002	1.05 (0.56-1.55) p = 0.000	0.85 (0.37-1.33) p = 0.001
FreeSurfer	**% Decrease 2008 to 2011**	0.42 (1.90) -5.96-2.51	1.75 (1.62) -2.15-4.86	2.08 (1.54) -0.97-4.98	3.06 (1.70) -0.16-5.95	3.46 (1.80) 0.13-6.30
	Significant difference	-	1.13 (0.06-2.21) p = 0.039	0.87 (0.37-1.37) p = 0.001	0.85 (0.49-1.22) p = 0.000	0.64 (0.31-0.97) p = 0.000

3.3.3.3.2. Cortical GM volume

The same analysis was conducted in CGM showing that CGM change was inconsistent across tools (Table 3.14 and Figure 3.16). Except for SPM8 New Segment, all tools showed significantly greater change in HD1 and HD2 compared to controls, SPM8 New Segment only showed greater change in HD1. MALP-EM, FAST and FreeSurfer all showed greater change in PreHD-B than controls, and only MALP-EM and FreeSurfer showed greater change in PreHD-B than controls.

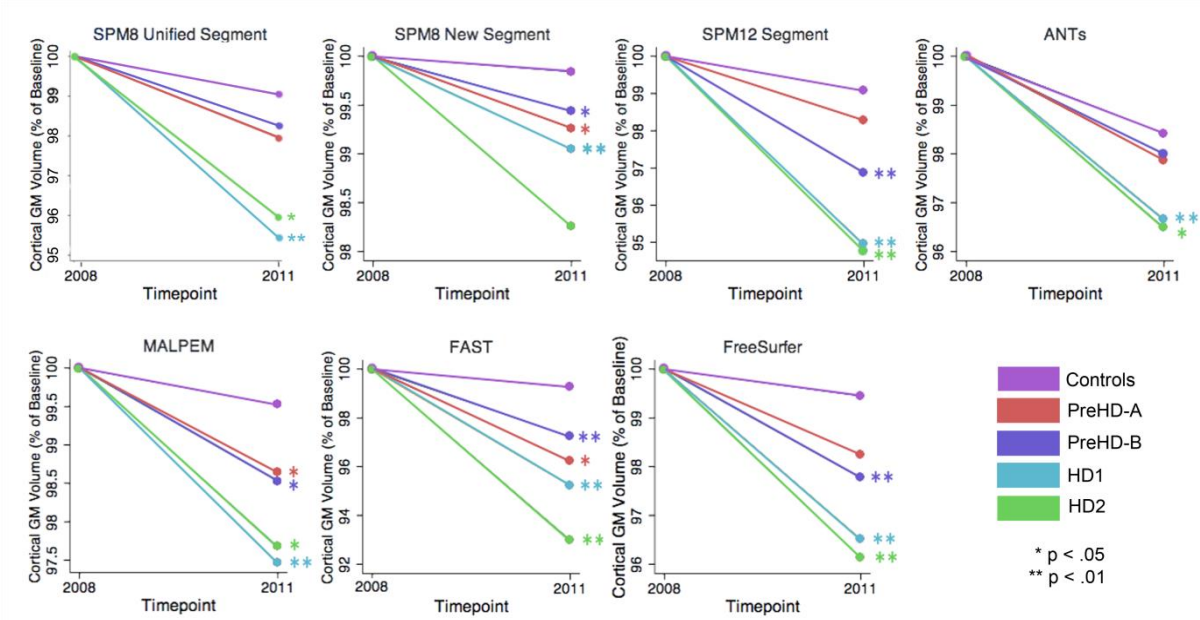


Figure 3.16. Mean values for all tools and groups showing 2011 volume as a percentage of baseline volume in CGM. Significant difference in longitudinal change relative to controls after controlling for age, sex and site are represented by * p<.05, **p<.01.

Table 3.14. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in cortical GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified Segment	% Decrease 2008 to 2011	.94 (3.06) -5.17-8.55	2.05 (2.17) -2.37-5.99	1.73 (3.71) -7.96-8.46	4.55 (4.11) -7.45-10.73	4.03 (2.30) -0.71-7.24
	Significant difference		.89 (-.82-2.59) p = 0.309	.40 (0.63-1.42) p = 0.447	1.26 (.56-1.97) p = 0.000	.55 (.06-1.03) p = 0.028
SPM8 New Segment	% Decrease 2008 to 2011	0.16 (0.93) -1.52-1.84	0.74 (0.90) -1.21-2.79	0.56 (0.62) -0.87-2.26	0.95 (0.74) -0.29-2.60	1.74 (3.22) -0.82-14.87
	Significant difference		-0.56 (-1.07--0.06) p = 0.030	-0.20 (-0.39-0.00) p = 0.046	-0.31 (-0.46--0.15) p = 0.000	-0.29 (-0.65-0.07) p = 0.114
SPM12 Segment	% Decrease 2008 to 2011	0.93 (2.35) -4.68-4.25	1.73 (3.87) -7.83-7.98	3.14 (1.71) 0.58-6.16	5.06 (2.39) 1.17-9.06	5.24 (2.61) 1.30-10.43
	Significant difference		-0.50 (-2.42-1.42) p = 0.608	-1.13 (-1.72--0.53) p = 0.000	-1.38 (-1.86--0.91) p = 0.000	-0.83 (-1.26--0.40) p = 0.000
ANTs	% Decrease 2008 to 2011	1.58 (2.75) -1.25-11.81	2.13 (2.55) -2.21-7.71	2.01 (1.90) -2.00-5.84	3.34 (1.79) 0.26-6.61	3.51 (3.95) -0.81-17.69
	Significant difference		-0.81 (-2.44-0.83) p = 0.335	-0.27 (-0.95-0.42) p = 0.446	-0.64 (-1.09--0.18) p = 0.006	-0.71 (-1.28--0.13) p = 0.017
MALP-EM	% Decrease 2008 to 2011	0.48 (1.24) -3.13-2.49	1.36 (1.26) -1.13-4.08	1.47 (1.24) -0.73-3.82	2.54 (2.69) -3.65-10.92	2.32 (2.10) -1.37-9.09
	Significant difference		-0.78 (-1.55--0.01) p = 0.047	-0.45 (-0.80--0.09) p = 0.013	-0.66 (-1.05--0.26) p = 0.001	-0.33 (-0.62--0.04) p = 0.027
FAST	% Decrease 2008 to 2011	0.74 (3.41) -6.63-8.33	3.77 (4.51) -2.62-12.54	2.77 (3.55) -0.36-15.00	4.78 (4.88) -3.85-14.14	7.00 (6.34) -0.09-20.82
	Significant difference		-1.96 (-3.48--0.45) p = 0.011	-1.12 (-1.83--0.41) p = 0.002	-1.09 (-1.59--0.59) p = 0.000	-0.85 (-1.37--0.33) p = 0.001
FreeSurfer	% Decrease 2008 to 2011	0.55 (1.49) -3.32-2.26	1.75 (1.77) -1.94-6.24	2.23 (1.86) -1.37-5.69	3.49 (2.04) -0.50-6.51	3.87 (2.04) -0.13-7.69
	Significant difference		-0.98 (-1.98-0.03) p = 0.057	-0.85 (-1.33--0.37) p = 0.000	-0.93 (-1.28--0.58) p = 0.000	-0.70 (-1.00--0.40) p = 0.000

3.3.3.3.3. Lobular GM volume

Longitudinal change within the lobes was variable for all groups with the parietal and occipital lobes showing the most consistent patterns of group differences across most techniques.

3.3.3.3.3.1. Frontal lobe volume

Table 3.15 shows the longitudinal change in the frontal lobe. SPM8 New Segment and ANTs showed no significantly different change in any group compared to controls. MALP-EM and SPM8 Unified Segment only detected significant change in HD1 compared to controls, and SPM12 detected change in PreHD-B, HD1 and HD2 compared to controls. Both FSL FAST and FreeSurfer found significant differences in longitudinal change in all HD groups.

3.3.3.3.3.2. Temporal lobe volume

For the temporal lobe, SPM8 New Segment, MALP-EM and ANTs showed no significant differences in volumetric change. FSL FAST and SPM8 Unified segment found significant differences in HD1 compared to controls, and FreeSurfer and SPM12 found differences in HD1 and HD2 compared to controls (Table 3.16).

3.3.3.3.3.3. Parietal lobe volume

Change in the parietal lobe was more widely detected across groups, with all techniques except SPM8 Unified Segment showing significantly greater volume reduction in HD1, HD2 and PreHD-B compared to controls, which only showed a difference between HD1 and controls. SPM8 New Segment, MALP-EM, FAST and FreeSurfer detected significantly greater change over time in all HD groups (Table 3.17).

3.3.3.3.3.4. Occipital Lobe Volume

Occipital lobe change was again widespread across all tools in HD groups compared to controls (Table 3.18). SPM8 New Segment, MALP-EM, FSL FAST and FreeSurfer found significantly greater change in all HD groups compared to controls. SPM12 and ANTs found significantly greater change in all PreHD-B, HD1 and HD2 compared to controls and SPM8 Unified Segment found greater change in HD1 and HD2.

3.3.3.3.3.5. Insula Volume

Finally, within the insula SPM8 New Segment and MALP-EM found no differences in change between any group and controls, SPM8 Unified Segment, ANTs, FAST and FreeSurfer found greater change in HD1 and HD2 than controls, and SPM12 found greater change in PreHD-B, HD1 and HD2 (Table 3.19).

Table 3.15. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in frontal lobe GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified Segment	**% Decrease 2008 to 2011**	1.24 (3.38) -5.62-8.99	2.01 (2.51) -3.26-6.17	1.82 (3.62) -8.46-8.73	4.21 (4.37) -8.55-11.33	4.05 (2.29) -1.57-7.31
	Significant difference		0.30 (-1.53-2.12) p = 0.751	0.23 (-0.81-1.27) p = 0.662	1.10 (0.34-1.85) p = 0.004	0.38 (-0.12-0.87) p = 0.137
SPM8 New Segment	**% Decrease 2008 to 2011**	0.44 (1.50) -1.75-3.92	0.83 (1.13) -2.16-3.31	0.73 (1.03) -0.85-2.82	0.75 (0.90) -0.94-2.41	1.90 (3.11) -0.32-14.45
	Significant difference		0.32 (-0.35-0.98) p = 0.349	0.12 (-0.22-0.46) p = 0.488	0.18 (-0.04-0.40) p = 0.115	0.23 (-0.16-0.61) p = 0.250
SPM12 Segment	**% Decrease 2008 to 2011**	1.15 (2.51) -5.25-4.17	1.69 (4.34) -9.59-8.19	3.22 (2.03) 0.26-6.40	4.72 (2.59) -0.16-9.58	5.12 (2.64) 0.87-11.27
	Significant difference	-	0.22 (-1.84-2.27) p = 0.838	1.03 (0.37-1.69) p = 0.002	1.23 (0.73-1.73) p = 0.000	0.73 (0.29-1.17) p = 0.001
ANTs Atropos	**% Decrease 2008 to 2011**	2.23 (5.15) -2.41-22.35	2.45 (2.41) -1.78-8.09	1.83 (2.26) -2.78-5.97	3.00 (1.95) -0.86-6.03	3.46 (3.86) -0.66-17.44
	Significant difference	-	0.39 (-2.07-2.85) p = 0.756	-0.15 (-1.32-1.02) p = 0.799	0.32 (-0.46-1.10) p = 0.423	0.56 (-0.25-1.37) p = 0.173
MALP-EM	**% Decrease 2008 to 2011**	0.91 (1.98) -4.01-3.70	1.81 (1.38) -0.36-4.33	1.67 (1.71) -1.03-4.92	2.56 (3.01) -3.95-12.09	2.57 (1.91) 0.40-8.28
	Significant difference	-	0.76 (-0.25-1.77) p = 0.141	0.28 (-0.24-0.79) p = 0.293	0.54 (0.08-1.01) p = 0.022	0.23 (-0.11-0.56) p = 0.181
FSL FAST	**% Decrease 2008 to 2011**	0.71 (3.59) -5.53-8.23	3.62 (5.60) -6.95-18.21	2.52 (4.06) -4.16-13.94	3.60 (5.91) -7.84-12.75	7.74 (9.14) -3.01-32.70
	Significant difference	-	1.70 (0.00-3.39) p = 0.050	0.94 (0.19-1.68) p = 0.014	0.77 (0.20-1.34) p = 0.008	0.96 (0.21-1.71) p = 0.012
FreeSurfer	**% Decrease 2008 to 2011**	0.55 (1.68) -1.68-3.49	2.11 (1.84) -1.37-6.96	2.05 (2.28) -1.59-6.33	3.14 (2.43) -1.39-7.06	3.41 (2.00) 0.26-6.45
	Significant difference		1.35 (0.30-2.41) p = 0.012	.71 (0.17-1.24) p=.010	.86 (.48-1.23) p = .000	.61 (.31-.91) p = .000

Table 3.16. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in temporal lobe GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified Segment	**% Decrease 2008 to 2011**	1.47 (3.00) -5.72-7.56	1.92 (1.93) -1.79-4.96	1.27 (4.02) -10.16-7.24	4.27 (3.82) -6.00-11.03	3.82 (2.42) -0.65-6.72
	Significant difference	-	0.46 (-1.12-2.03) p = 0.569	-0.07 (-1.06-0.91) p = 0.887	0.90 (0.25-1.55) p = 0.006	0.36 (-0.10-0.83) p = 0.122
SPM8 New Segment	**% Decrease 2008 to 2011**	0.75 (1.53) -1.63-3.97	0.44 (1.29) -1.67-3.40	0.39 (0.99) -2.22-1.71	0.70 (1.21) -1.97-4.11	1.48 (3.90) -3.90-16.23
	Significant difference	-	-0.16 (-0.95-0.63) p = 0.692	-0.16 (-0.48-0.16) p = 0.330	-0.06 (-0.31-0.18) p = 0.617	0.05 (-0.40-0.49) p = 0.839
SPM12 Segment	**% Decrease 2008 to 2011**	1.41 (2.35) -5.56-5.02	1.49 (3.21) -5.70-6.56	2.49 (2.03) -1.48-6.04	4.31 (2.58) 0.70-10.08	4.80 (3.20) -0.51-10.75
	Significant difference	-	-0.13 (-1.81-1.55) p = 0.882	0.56 (-0.04-1.16) p = 0.067	0.88 (0.43-1.32) p = 0.000	0.60 (0.16-1.03) p = 0.008
ANTs Atropos	**% Decrease 2008 to 2011**	2.63 (3.50) -0.91-15.76	2.11 (2.80) -2.32-8.83	1.64 (1.98) -1.96-5.58	3.06 (1.81) -0.02-6.48	3.43 (4.62) -2.50-19.77
	Significant difference	-	-0.11 (-1.95-1.72) p = 0.904	-0.42 (-1.18-0.35) p = 0.284	0.10 (-0.44-0.64) p = 0.721	0.47 (-0.18-1.12) p = 0.160
MALP-EM	**% Decrease 2008 to 2011**	1.26 (1.54) -3.59-4.26	1.29 (1.75) -1.64-5.62	1.28 (1.54) -1.49-5.11	2.34 (3.95) -8.26-14.21	2.19 (3.20) -4.65-12.75
	Significant difference	-	-0.03 (-1.02-0.95) p = 0.950	0.00 (-0.40-0.40) p = 0.998	0.19 (-0.38-0.77) p = 0.512	0.04 (-0.35-0.43) p = 0.829
FSL FAST	**% Decrease 2008 to 2011**	1.47 (3.32) -5.72-7.87	3.72 (4.27) -3.02-10.72	2.13 (2.79) -1.40-11.11	4.38 (4.92) -3.65-17.89	5.13 (3.90) -1.31-14.52
	Significant difference	-	1.33 (-0.41-3.07) p = 0.135	0.43 (-0.30-1.15) p = 0.248	0.65 (0.05-1.25) p = 0.035	0.34 (-0.09-0.78) p = 0.124
FreeSurfer	**% Decrease 2008 to 2011**	1.18 (2.43) - 3.25-7.54	1.49 (2.55) -4.84-6.39	1.58 (1.92) -1.87-5.11	3.42 (1.73) 1.19-7.38	4.04 (2.45) -2.29-7.79
	Significant difference		.23 (-1.26-1.72) p = 0.762	.28 (-.32-.87) p = 0.364	.71 (.71-1.12) p = 0.001	.63 (.24-1.02) p = 0.002

Table 3.17. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in parietal lobe GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified Segment	**% Decrease 2008 to 2011**	0.44 (3.19) -4.34-8.36	2.24 (2.06) -2.31-5.63	2.13 (3.59) -6.23-9.01	4.60 (4.06) -7.72-10.10	3.81 (2.69) -1.66-7.53
	Significant difference	-	1.62 (-0.06-3.31) p = 0.059	0.85 (-0.21-1.90) p = 0.115	1.48 (0.76-2.19) p = 0.000	1.62 (-0.06-3.31) p = 0.059
SPM8 New Segment	**% Decrease 2008 to 2011**	-0.37 (1.67) -3.77-2.14	0.81 (0.97) -1.07-2.37	0.60 (0.84) -1.24-1.86	1.05 (1.13) -1.31-2.93	1.56 (2.29) 0.03-10.58
	Significant difference	-	1.05 (0.36-1.74) p = 0.003	0.45 (0.11-0.78) p = 0.009	0.55 (0.31-0.80) p = 0.000	0.36 (0.04-0.67) p = 0.027
SPM12 Segment	**% Decrease 2008 to 2011**	0.51 (3.00) -7.75-4.38	1.70 (3.61) -8.07-7.88	3.33 (1.48) 0.91-6.51	5.22 (2.28) 1.25-9.39	4.95 (2.91) 0.59-9.92
	Significant difference	-	0.83 (-1.15-2.81) p = 0.412	1.42 (0.75-2.10) p = 0.000	1.61 (1.06-2.15) p = 0.000	0.80 (0.30-1.30) p = 0.002
ANTs Atropos	**% Decrease 2008 to 2011**	0.47 (1.94) -2.63-5.38	1.98 (2.68) -4.78-7.28	2.46 (2.28) -1.66-8.58	3.54 (2.10) -0.37-6.95	3.30 (3.24) -2.03-12.84
	Significant difference	-	1.29 (-0.18-2.75) p = 0.085	0.96 (0.35-1.57) p = 0.002	1.11 (0.74-1.48) p = 0.000	0.68 (0.22-1.13) p = 0.003
MALP-EM	**% Decrease 2008 to 2011**	0.07 (1.57) -3.95-2.40	0.97 (1.00) -1.03-3.07	1.39 (1.22) -0.65-3.56	2.23 (1.68) -0.59-7.16	1.94 (1.55) -0.60-6.01
	Significant difference	-	0.77 (0.05-1.48) p = 0.036	0.60 (0.21-0.99) p = 0.003	0.76 (0.48-1.03) p = 0.000	0.34 (0.08-0.60) p = 0.011
FSL FAST	**% Decrease 2008 to 2011**	0.09 (3.41) -9.04-5.46	3.27 (3.93) -2.64-11.00	3.15 (2.96) -0.48-11.11	4.24 (3.65) -4.03-11.29	5.30 (4.42) 0.72-17.44
	Significant difference	-	2.23 (0.54-3.92) p = 0.010	1.54 (0.77-2.32) p = 0.000	1.25 (0.72-1.78) p = 0.000	0.80 (0.24-1.36) p = 0.005
FreeSurfer	**% Decrease 2008 to 2011**	0.45 (1.41) -2.55-2.50	1.45 (1.94) -2.51-5.97	2.96 (2.14) -1.60-7.27	3.89 (2.54) 0.33-9.60	3.86 (2.38) -0.78-9.35
	Significant difference		.60 (-.40-1.61) p = 0.239	1.21 (.70-1.72) p = 0.000	1.06 (.67-1.44) p = 0.000	.65 (0.35-0.94) p = 0.000

Table 3.18. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in occipital lobe GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified Segment	% Decrease 2008 to 2011	0.35 (4.42) -11.14-9.57	2.18 (3.12) -4.03-8.78	1.88 (4.26) -8.03-9.34	6.06 (4.87) -7.32-15.50	5.01 (3.26) -0.43-9.77
	Significant difference	-	1.81 (-0.79-4.41) p = 0.173	0.90 (-0.46-2.26) p = 0.197	1.95 (1.01-2.89) p = 0.000	1.31 (0.48-2.14) p = 0.002
SPM8 New Segment	% Decrease 2008 to 2011	-0.66 (2.13) -7.72-1.27	0.96 (1.95) -3.02-4.27	0.39 (1.60) -1.81-3.56	1.78 (1.22) -0.16-4.12	2.40 (4.68) -2.66-20.47
	Significant difference	-	1.85 (0.60-3.11) p = 0.004	0.56 (0.01-1.11) p = 0.047	0.83 (0.47-1.19) p = 0.000	0.76 (0.18-1.34) p = 0.010
SPM12 Segment	% Decrease 2008 to 2011	0.58 (3.47) -9.26-4.94	2.28 (5.04) -7.40-11.37	3.88 (2.72) -0.14-9.25	7.11 (3.24) 2.44-15.85	7.12 (3.32) 1.68-12.39
	Significant difference	-	1.59 (-1.11-4.30) p = 0.248	1.74 (0.83-2.66) p = 0.000	2.15 (1.47-2.83) p = 0.000	1.42 (0.82-2.03) p = 0.000
ANTs Atropos	% Decrease 2008 to 2011	0.35 (2.95) -8.51-6.80	1.90 (3.98) -3.01-8.94	2.79 (2.95) -2.22-8.53	4.64 (2.58) 0.59-10.90	4.65 (5.08) -2.34-22.64
	Significant difference	-	1.69 (-0.44-3.82) p = 0.119	1.32 (0.46-2.18) p = 0.003	1.45 (0.91-1.98) p = 0.000	1.21 (0.51-1.91) p = 0.001
MALP-EM	% Decrease 2008 to 2011	-0.73 (2.10) -7.74-1.90	1.10 (2.05) -3.06-4.50	1.52 (2.16) -4.24-5.23	3.57 (2.56) -0.53-9.38	2.99 (3.04) -3.59-11.72
	Significant difference	-	1.93 (0.66-3.19) p = 0.003	1.17 (0.54-1.80) p = 0.000	1.36 (0.90-1.81) p = 0.000	0.92 (0.48-1.36) p = 0.000
FSL FAST	% Decrease 2008 to 2011	0.63 (6.00) -18.17-13.82	5.17 (6.56) -3.67-22.14	4.22 (6.99) -0.25-31.20	9.45 (10.13) 0.13-44.52	12.09 (12.81) -1.89-37.55
	Significant difference	-	3.47 (0.52-6.41) p = 0.021	2.13 (0.71-3.55) p = 0.003	2.26 (0.99-3.53) p = 0.000	1.82 (0.75-2.88) p = 0.001
FreeSurfer	% Decrease 2008 to 2011	-0.08 (2.84) -10.87-2.81	1.68 (2.54) -3.08-6.66	3.06 (2.99) -3.23-7.75	4.71 (2.49) -1.62-8.27	4.96 (3.43) 0.17-11.70
	Significant difference		1.74 (0.12-3.37) p = 0.036	1.71 (0.88-2.54) p = 0.000	1.57 (1.02-2.11) p = 0.000	1.28 (0.73-1.84) p = 0.000

Table 3.19. Mean % change (difference between 2011 and baseline volume as a percentage of baseline volume), standard deviation and ranges for all tools and groups in insula lobe GM. Positive values represent volumetric decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
SPM8 Unified	**% Decrease 2008 to 2011**	0.28 (3.39) -9.01-5.73	1.50 (2.35) –3.92-5.17	1.68 (2.59) -3.62-5.56	3.97 (3.19) -4.54-8.96	4.11 (3.36) -2.20-13.00
	Significant difference	-	0.75 (-1.14-2.65) p = 0.438	0.74 (-0.19-1.67) p = 0.118	1.22 (0.58-1.86) p = 0.000	1.07 (0.44-1.70) p = 0.001
SPM8 New	**% Decrease 2008 to 2011**	-0.77 (2.95) -11.79-2.19	0.17 (1.20) -2.01-1.94	0.54 (1.69) -2.97-4.18	0.23 (1.24) -3.30-2.12	1.35 (5.08) -3.15-21.44
	Significant difference	-	0.86 (-0.51-2.23) p = 0.218	0.65 (-0.05-1.35) p = 0.070	0.32 (-0.12-0.77) p = 0.154	0.62 (-0.09-1.33) p = 0.086
SPM12	**% Decrease 2008 to 2011**	-0.13 (3.01) -9.85-3.30	1.55 (3.02) -5.67-7.31	2.64 (1.84) 0.00-6.52	4.49 (2.42) -0.47-8.85	5.27 (4.37) 0.22-16.80
	Significant difference	-	1.39 (-0.45-3.24) p = 0.138	1.41 (0.68-2.14) p = 0.000	1.52 (0.98-2.06) p = 0.000	1.32 (0.65-1.99) p = 0.000
Atropos	**% Decrease 2008 to 2011**	-0.23 (3.71) -13.19-4.79	0.25 (2.10) -4.82-3.59	0.51 (1.66) -3.63-3.92	1.85 (1.73) -0.97-4.60	3.16 (5.20) -1.54-21.45
	Significant difference	-	0.41 (-1.33-2.14) p = 0.645	0.40 (-0.44-1.23) p = 0.351	0.69 (0.18-1.20) p = 0.008	1.20 (0.42-1.98) p = 0.003
MALP-EM	**% Decrease 2008 to 2011**	-1.65 (6.72) -23.50-10.04	1.07 (2.94) -1.91-10.41	1.18 (3.22) -2.53-12.26	1.07 (4.21) -8.27-8.34	1.47 (3.70) -3.90-11.62
	Significant difference	-	2.96 (-0.16-6.08) p = 0.063	1.30 (-0.27-2.86) p = 0.104	0.72 (-0.37-1.82) p = 0.196	0.77 (-0.24-1.78) p = 0.136
FAST	**% Decrease 2008 to 2011**	-0.72 (5.10) -19.77-4.00	1.99 (2.83) -2.61-8.31	1.71 (2.12) -1.16-8.53	2.90 (1.91) -1.34-5.73	3.49 (2.41) 0.59-9.00
	Significant difference	-	1.97 (-0.11-4.04) p = 0.063	1.19 (0.13-2.25) p = 0.028	1.13 (0.44-1.81) p = 0.001	0.93 (0.21-1.64) p = 0.011
FreeSurfer	**% Decrease 2008 to 2011**	-0.69 (4.26) -16.16-3.23	1.29 (2.49) -1.88-7.00	1.13 (2.47) -4.12-5.36	1.74 (2.42) -1.79-6.90	3.43 (3.51) -3.24-14.01
	Significant difference		1.56 (-0.45-3.56) p = 0.128	0.84 (-0.18-1.87) p = 0.106	0.72 (0.03-1.41) p = 0.041	0.95 (0.27-1.63) p = 0.006

3.4. Discussion

The detailed analysis of GM segmentation techniques undertaken in this research has provided useful insight into the measurement of grey matter volume within HD, especially for measurement of the CGM. The combined qualitative and quantitative analyses conducted on two different datasets provided great depth of information that allows informed choices to be made on the best techniques for use on the TRACK-HD cohort, but also offers lessons for researchers working on other cohorts. The results demonstrate that most tools show good

reliability when assessed via quantitative measures, but that visual QC highlights variability between tools. Furthermore, the tools show different patterns of results when using them to measure longitudinal change, supporting previous work that highlights the impact of different methods on the conclusions drawn in neuroimaging (Katuwal et al. 2016).

An initial finding that warrants discussion is that the user input required for each tool varies widely. Some tools required processing steps prior to the segmentation (e.g. ANTs), whereas others did not (e.g. SPM, FreeSurfer). In addition, some required the ability to optimise the pipeline for different datasets by creation of a template or probability priors. While some users may prefer greater flexibility and customisation options from a tool, others may want a more simple application with limited options. Based on the experience of running seven tools on both BrainWeb and TRACK-HD datasets, all tools except ANTs were user-friendly and understandable. Despite having experience in some coding and an understanding of concepts related to the processing of MRI data, ANTs was still difficult to optimise and obtain satisfactory results from. While there is evidence that ANTs is reliable and accurate (Tustison et al. 2014; Avants, Tustison, Wu, et al. 2011; Schwarz et al. 2016), these results were not replicated and it is assumed that this was because of the difficulty experienced when trying to apply the tool to the current datasets. Whilst satisfactory results were achieved, it involved a lot of experimentation, and the results were inferior to what was expected given the previously reported performance of ANTs. Interestingly, this replicates a recent study which also found poorer performance on ANTs than anticipated (Righart et al. 2017) For a novice user with little understanding of the multitude of options required to run tools such as ANTs, optimisation can be a long process of trial and error rather than making informed choices about the best options to select for each dataset. Most other tools provided default options that were largely successful.

Another difference between tools that is important to recognise is that many define GM regions differently; some include subcortical structures, cerebellar GM and brainstem, and some do not. In addition, some only provide whole-brain results and some divide the GM into large or small regions. The use of PVE GM versus discrete GM volumes is inconsistent between the tools, with most tools outputting both partial volume corrected maps and discrete tissue maps (FSL, ANTs, MALP-EM) but some outputting more complex volumetric maps (FreeSurfer). Some software also output pre-calculated volumes for the segmentations and regions in text files. Table 3.20 shows a summary of these factors for the software included in this analysis.

Table 3.20. A summary of some characteristics of the tools included in the current study.

	SPM8 Unified Segment	SPM8 New Segment	SPM12 Segment	ANTs	MALP-EM	FAST	FreeSurfer
Total GM volume?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Regional Volume as default?	No	No	No	No (additional steps required)	Yes	No	Yes
Partial Volume Maps?	Yes	Yes	Yes	Yes	Yes	Yes	No
Discrete Volume Maps?	No	No	No	Yes	Yes	Yes	No
Volumes automatically output in text file?	No	No	No	No	Yes	No	Yes

Qualitative analysis was done prior to quantitative analysis for both datasets. For the BrainWeb data initial examination of the output enabled greater understanding of the performance of each technique in a near perfect dataset. While the quantitative analysis can measure the performance of a number of metrics, only visual QC can detect subtle and persistent factors in the regions, such as over- or underestimation of the GM. Whilst reviewing the segmentations, the GT for each scan was examined and it became clear that GT regions are slightly inaccurate. The segmentations included dura in a moderate number of scans, and this should be considered whilst drawing conclusions from this analysis. Due to the errors associated with the BrainWeb GT, a total overlap between the GT and segmentation tools should not be expected as some of the tools may more accurately delineate the cortical surface.

SPM8 Unified Segment was the first tool to be examined. The border between WM and GM was gradual, with a large number of voxels classified as having some proportion of GM although upon visual inspection they were WM voxels. In the BrainWeb dataset the GM/CSF boundary was tight, although this was not an issue in the TRACK-HD cohort. While the software performed well in the artificial BrainWeb data, in the TRACK-HD cohort SPM8 Unified Segment poorly segmented the occipital GM and temporal poles, with these areas frequently being overestimated. There was also occasional underestimation of the frontal lobes. This indicates that on a real dataset with increased noise and movement the software suffers from increased errors in segmentation. SPM8 Unified Segment also had the highest rate of

segmentation failures, with 11 initial fails in the TRACK-HD data (3.67% of the dataset), reducing to 6 total fails (2% of the dataset) after re-aligning and re-running the data in the TRACK-HD cohort.

In both the BrainWeb and TRACK-HD segmentations SPM8 New Segment introduced significant error through the incorrect classification of voxels in the dura, skull and outside the skull as GM. While the PVE for these voxels were low, this error was persistent and if these segmentations were used for a whole study it is likely that the results would be less sensitive to group differences or longitudinal change due to the incorrect classification of non-brain tissue. In addition, SPM8 New Segment also showed problems with overestimation of the occipital and temporal lobes.

SPM12 showed an improvement on both earlier versions of SPM. The boundaries between GM/WM were clearer and for the TRACK-HD cohort the delineation in occipital and temporal regions was improved, although there was still minor overestimation and spillage into the dura in these regions. Based on visual inspection, SPM12 output the most accurate version of SPM.

ANTs was the most variable software, likely due to the increased user input required to process scans using this software. While this allows for more flexibility and greater applications of the software, it can also result in variable results due to minor differences in steps such as template creation or registration. For the BrainWeb dataset ANTs performed well, showing good delineation of the sulci and no persistent errors. Performance in the TRACK-HD dataset appeared to be highly dependent on brain size, with small brains including more dura. For this cohort a standard group template was created resulting in a template that was too large for the HD participants, but it is possible that if a template was created for each HD group the results could be improved. Similarly to SPM, the most common regions for errors were the occipital and temporal lobes – with overestimation the most persistent error, although the amount of overestimation varied largely.

MALP-EM was the most visually consistent software both within and between the two cohorts. In the BrainWeb dataset it suffered from minor underestimation of the temporal regions in some cases, but performed well otherwise. In the TRACK-HD cohort the consistent errors associated with occipital and temporal segmentation in other techniques appeared much more infrequently in MALP-EM regions. There were minor examples of dura inclusion or exclusion in some scans around the superior boundaries.

FAST performed very differently on the BrainWeb and TRACK-HD cohorts. The segmentations on the BrainWeb dataset were very good, with good delineation of the surface

and only minor spillage in occipital and temporal lobes. Comparatively, FSLs FAST performed very poorly on the TRACK-HD cohort. This was mostly due to the failure of the initial brain extraction, BET, to successfully detect and extract the brains in this dataset despite using an optimised procedure. Large chunks of the frontal and occipital lobes were often omitted from the segmentations, and spillage into other brain regions was also a problem for these segmentations. The contrast between these two scenarios highlight the importance of testing all tools before applying them to different datasets, since software which performs very well in one dataset may perform poorly in another due to the large variability associated with MRI scanning parameters and different cohorts.

FreeSurfer performed very poorly on the BrainWeb data. Large regions of the GM were omitted from almost every segmentation, especially in occipital, frontal and temporal lobes. While FreeSurfer segmentations for the TRACK-HD data were an improvement, the regions were generally tight, suggesting overall underestimation of the GM, and there were a number of segmentation errors, with the most common error being overestimation of the occipital and superior regions of the brain.

The qualitative results showed large variability in performance, with SPM12 Segment and MALP-EM appearing as the most consistent and accurate methods after visual QC. In addition, with further optimisation of the group template and segmentation procedure on the TRACK-HD cohort it is likely that the results from ANTs could be improved to reduce errors associated with inclusion of extra dura. Of these tools, MALP-EM performed the most consistently in the TRACK-HD cohort with less overestimation of occipital and temporal regions, and only rare cases of the GM region extending into dura or cortical volume excluded from the region. Throughout this process, over 2000 segmentations were reviewed and it quickly became apparent how vital the visual QC process is. A number of scans failed processing initially, either due to user error such as the incorrect parameters being specified or due to failure in the software. When examining the volumetric values output from failed segmentations it was not always obvious from these values that the segmentation was very poor. This supports one previous study in which outliers were used to select the scans that require visual QC. The authors found that this technique resulted in a large number of segmentation errors being missed, and concluded that all data should undergo visual QC (Waters et al. 2017). Our results confirm this, as often when the volumetric values of a failed segmentation were compared to other values for the same region, they did not appear to be quantitative outliers. When performing a longitudinal or regional analysis, the inclusion of failed data would have a large impact of the results. While only a small number of scans did

completely fail processing in the current study, there were a greater proportion of poor segmentations – especially for FAST and FreeSurfer and particularly in the occipital and temporal lobes. When performing analysis on a small cohort or trying to detect subtle brain changes within a group these poor segmentations would likely have a large impact on the sensitivity of tools. In these instances, more stringent inclusion criteria would be beneficial.

The quantitative results from the BrainWeb analysis provide evidence on the accuracy of these tools. Initially, the BrainWeb cohort was used to compare both partial volume maps and discrete volume maps to the GT. For all segmentation techniques that provided partial volume maps, the values were much closer to those of the GT, and mean differences were generally lower, suggesting a higher level of accuracy. This is consistent with previous reports (Tohka 2014). While the dice scores for most measures were slightly lower for partial volume maps, it is likely that this is due to the voxel-wise calculation of dice scores in this study. For discrete volumes the voxel values were either 1 or 0, corresponding to GM or not GM, whereas for partial volume maps the voxels could vary between 0 and 1 and thus a lower level of overlap in the partial volume scores is unsurprising.

Total and cortical GM results for the BrainWeb analysis suggested that while there was good correspondence between BrainWeb values and the output volumes for the different tools, there were still large differences in the raw volumes. This was especially apparent for FreeSurfer, which had a much greater difference than any other method for cortical GM, significantly under-estimating cortical GM. While a complete overlap was not expected due to the errors in BrainWeb segmentations previously mentioned, the large divergence between BrainWeb and FreeSurfer is concerning, and supports the poor results of the qualitative analysis examining FreeSurfer. The other tools overall performed well on the BrainWeb data, with small, albeit significant, differences between the GT volumes and segmented volumes. The ICC values from all tools provided evidence that there is good overlap despite the significant differences in raw volumes.

The lobular results provide more specific information on the performance of these tools, especially FreeSurfer. The values for most tools were close to the expected values based on the GT. Differences between the raw GT and lobular volumes for most tools were around 3-4%, and although these differences were significant in a number of cases, as measured by paired t-tests, the overlap measures indicated voxel-wise correspondence of around 0.90. This demonstrates very high quantitative overlap with the GT. FreeSurfer showed large differences to the lobular GT volumes for all lobes. While it is likely that this is, in part, due to the fact that the automatically calculated regions were used for the FreeSurfer lobular analysis (see section

3.2.4), it also demonstrates the poor performance of FreeSurfer on the BrainWeb data as verified by the visual QC. Unfortunately, it is impossible to calculate voxel-wise overlap for FreeSurfer as voxel-wise maps are not output during processing. However, the visual and quantitative results provide enough evidence to demonstrate very poor performance of FreeSurfer on the BrainWeb data.

The wide variation in the quality of segmentation performance on the TRACK-HD data is further demonstrated by the quantitative results, which showed large differences in segmentation across tools for raw total, cortical and lobular GM volumes. Longitudinal analysis demonstrated that while the pattern of total and cortical GM change was similar across tools, when GM change in HD participants was statistically compared to GM change in controls, the tools detected differing degrees of change. In addition, for lobular volumes the tools showed different patterns of change in a number of regions. Despite the variability in raw volumes and sensitivity to change, all tools generally showed high reliability across groups and regions, as measured by ICC and repeatability metrics, and extracted volumes were typically highly correlated between different tools.

For controls and HD participants, GM regions derived using MALP-EM were consistently larger than those for other tools, likely due to the regions having higher overall probabilistic GM segmentation values. SPM also outputted comparatively large regions, with FSL FAST and FreeSurfer outputting the lowest; a discrepancy that has been previously noted (Katuwal et al. 2016). SPM regularly over-estimated both occipital and temporal lobe regions, this was particularly noticeable for larger brains and in earlier versions of SPM, indicating that SPM may over-estimate between-group differences when comparing healthy to atrophied brains. Low volumes output from FSL segmentations can be explained by regular underestimation of the cortical GM, partly resulting from poor brain extraction in this cohort. It is important to note that the poor performance of BET should not rule out the use of FSL in other cohorts, since it is known to perform well on other data (Smith 2002). Lower volumes output by FreeSurfer are possibly due to subtle underestimation of the GM occurring at the cortical boundary with CSF. While the partial volume included via calculation of FreeSurfer regions would account for some of these voxels, the regions remained tight along the outer boundary after accounting for this. ANTs showed average values that were in the middle of most tools, possibly due to the variable performance of this tool.

Although all tools demonstrated errors using the default pipelines, in this analysis no manual intervention was performed to improve the quality of segmentation due to the increased subjectivity involved in manual intervention. However, it is important to note that all

tools offer some opportunity to improve the issues described above via optimisation of the segmentation and/or manual intervention after segmentation. For example, FreeSurfer allows editing of the segmented region to improve regions with over- or under- segmentation, MALP-EM, FAST, SPM and ANTs can all be used with highly optimised masks to improve segmentation, or can be manually edited after segmentation.

Despite the lack of optimisation and manual editing it is reassuring to note good reliability for back-to-back scans across tools using ICC and a repeatability metric. These findings support the results of previous studies comparing SPM, FSL and FreeSurfer (Katuwal et al. 2016) and provide new results validating ANTs and MALP-EM. Most tools were also highly correlated, again supporting previous studies (Katuwal et al. 2016). Furthermore, there was no indication from the quantitative results that the tools have any difference in reliability between control participants and HD participants. Although correlations between the techniques tended to decrease with increasing disease progression, indicating some divergence of performance between the tools on brains with more advanced atrophy, particularly for SPM8 Unified Segment, which resulted in smaller volumes for HD participants and may indicate an under-estimation of GM in HD. This result is similar to a previous study, which indicated possible over-estimation of group differences in SPM (Callaert et al. 2014). Using optimised brain masks or performing manual editing may improve performance on atypical brains, but generally the tools performed as well on HD participants as they did on control participants.

Different techniques showed different longitudinal sensitivity to GM change in HD groups compared to controls, especially within the five lobes. For total and cortical GM, all techniques showed decreasing volume with increasing disease progression. While SPM8 Unified Segment, SPM12 and FAST showed the largest decreases in total GM volume over time, MALP-EM and FreeSurfer both showed significant change across all disease stages, possibly indicating greater sensitivity to small changes. In cortical GM, SPM8 Unified Segment, SPM12 and FAST showed large decreases in raw GM volume, but again MALP-EM and FreeSurfer showed statistical sensitivity to small changes. MALP-EM was developed for use in clinical cohorts, which could partly explain the sensitivity of MALP-EM in an HD cohort (Ledig et al. 2015).

The results of the longitudinal lobular analysis indicate large differences between the tools on each lobe, with particular divergence between tools in the frontal lobe, temporal lobe and insula. The results of this analysis emphasise the importance of good quality segmentations, with some significant results being driven by participants with very high rates

of change. For example, a significantly greater volumetric loss in PreHD-A compared to controls in the frontal lobe as measured by FSL FAST was possibly driven by a percentage loss in one participant of 18% between baseline and follow-up. Re-examination of the segmentations revealed underestimation of the frontal lobe in the follow-up time point rather than a true volumetric change. If manual editing had been performed this result would be accounted for, and a more biologically plausible rate of change measured. The impact of segmentation errors is reduced in large cohorts or with whole-brain analyses, but when examining regional change the quality of every slice of the segmentation is paramount. This result supports the work of (Iscan et al. 2015) who found that by including scans which have poorly delineated FreeSurfer regions in an analysis, the sample size required to detect a true effect increases. It is possible that the results shown in the parietal and occipital lobe are more uniform across tools as this is a region that is thought to undergo the most atrophy within cortical GM in pre-HD and early HD (Johnson et al. 2015; Rüb et al. 2015; Tabrizi et al. 2012; Wolf et al. 2014), and so statistical sensitivity to change in these regions is more robust when segmentation quality is sub-optimal and variability is high. In comparison, regions that may have slower rates of change in HD (e.g. frontal lobe) are more easily biased by poor segmentations and thus different techniques identify different patterns of atrophy. Interestingly, all techniques found a higher volume in PreHD-A than in controls. This result is consistent with TRACK-HD, and may indicate a compensatory mechanism in the pre-HD group However, the raw volumes are not corrected for TIV and so these findings should be further investigated. This study is underpowered to draw conclusions on the true nature of lobular progression of GM atrophy in HD, the results provide a useful demonstration on the importance of selecting a segmentation tool that performs well on a particular dataset.

3.4.1. Suggestions for the selection of a segmentation tool

The results of this study can inform the selection of a GM segmentation tool for use in the TRACK-HD cohort, but they can also be generalised to other clinical cohorts, particularly neurodegenerative diseases. All tools showed consistently high reliability when used with clinical data. However, there were a significant number of segmentation errors and while segmentation quality for each tool can be optimised, it is important to note variable results are likely depending on factors including scanner parameters, quality of data and researcher expertise. This analysis can help to provide a set of guidelines on how to select of the most appropriate segmentation tool in different types of datasets, as described here.

3.4.1.1. Which segmentation tool best answers your question?

It is important to consider which software tools contain features that are most appropriate for addressing your question or hypothesis. For example, if regional GM volume is the main focus, considering software that includes the option to perform regional analysis and atlas optimisation can potentially reduce errors otherwise introduced when applying a mask to your cohort. Some tools (for example, SPM, ANTs, FSL, MALP-EM) are suitable for extracting volumes from customised regions or atlases applied to volumetric maps, while others, such as FreeSurfer, recommend using default atlases. Furthermore, the intended use of other structural analyses such as CT or VBM should also be considered.

3.4.1.2. Are your knowledge, experience and resources sufficient to use the tool?

While some tools offer numerous options for customisation, this often requires an appropriate level of analysis and computing expertise. Without full understanding of the changes being applied to a pipeline, there is a risk of producing results that are not accurate, reliable or reproducible. Conversely, some tools provide little or no customisation, and thus are simple to apply yet may be limited in the range of appropriate applications. If available, use your chosen tool to analyse test data with the same MRI acquisition parameters as those in your study prior to beginning the analysis on a full cohort of participants. This will reduce delays due to errors or complexity of technique when applying it to a large cohort.

As an example, ANTs has previously been validated with impressive results for both registration and CT using default pipelines in a healthy cohort (Avants, Tustison, Song, et al. 2011; Tustison et al. 2014). However, due to the number of possible optimisations for a clinical cohort, such as creation of a study-specific template and priors, an inexperienced user may find using this tool challenging. When applying ANTs to the TRACK-HD cohort, the large variation in brain sizes within each patient group meant that when using the study-specific brain mask to extract the brain, dura was often erroneously classified as GM. Group-specific templates may have been more suitable here, but this would have required more time and expertise to create and optimise.

Another consideration is the variable nature of processing time necessary for each tool, which can range from 5 minutes per brain (SPM) to 24 hours (FreeSurfer). In addition, a number of registration or segmentation tools require high levels of processing power and may not run locally on desktop/laptop machines. Access to a high-powered computer cluster or just a single laptop with limited processing resources might determine which tool you choose. Financial costs also warrant consideration. While all tools examined for

this case study are freely available for academic purposes, SPM works on a MATLAB platform that is not freely available. In addition, some tools require a licence for use in industry settings and if developing methodology for later use within clinical trials, industry costs of these tools should be considered.

3.4.1.3. Which segmentation tools are most reliable?

A number of studies have previously demonstrated that some of the tools discussed in the current study have high reliability (Eggert et al. 2012; Klauschen et al. 2009). These findings were supported by current results in the TRACK-HD cohort. Current versions of all tools, including two not previously validated in this way (ANTs Atropos and MALP-EM), demonstrated high reliability. While repeatability was more variable, and lower in early-HD participants, the values were still very good for most tools. While reliability can be established for all tools in this and previous studies, accuracy is more difficult to determine – see 3.4.1.4.

3.4.1.4. Which segmentation tools are most accurate?

While phantom data can be used to examine the accuracy of volumetric measurement tools in healthy models, the results of phantom analyses do not always represent performance when applied to clinical cohorts. In HD, for example, where the pattern of neural change is not well understood it is challenging to define the tool that provides the greatest accuracy. Often, measures are examined in terms of their overlap to estimate accuracy – however since all tools showed consistent errors (e.g. over-segmentation in occipital and temporal regions) this could falsely inflate accuracy results. How do we determine which is the most accurate result? Visual QC was an important factor when assessing accuracy in this study. MALP-EM appeared to be the most visually accurate tool, with SPM12 also performing well despite some segmentation errors in temporal and occipital regions. By comparing the results of longitudinal change in this sample to other previously published values of longitudinal GM change in larger HD cohorts (measured by various techniques), it appears that a number of techniques over-estimate GM change in this small cohort. SPM8 New Segment, MALP-EM, ANTs and FreeSurfer produced values in line with previous studies suggesting that they may be able to detect accurate results from a small sample (Juan et al. 2016; Tabrizi et al. 2012). Since inconsistencies between volumetric neuroimaging tools are thought to result in contradictions within clinical neuroimaging research (Ashburner et al. 2016), it is imperative that the accuracy of each tool is visually examined within a cohort to ensure good performance of the tools on each dataset.

3.4.1.5. Should I perform visual QC of my data?

This is a necessary step in the processing of any data. All registrations, segmentations and masks in this study were visually checked. Errors in processing, complete segmentation failures or patterns in segmentation errors were only detected by viewing the data. While total failure of a tool to segment an image was rare, consistent errors in segmentation were common. In cases where segmentation did fail completely, the volumetric measurements for total GM and CGM were often not out of line with expected values and so may not be detected in a quantitative check of extreme values, but would not provide reliable data on pathology or longitudinal sensitivity to change. This replicates previous findings (Waters et al. 2017).

In this study, for cases where segmentation was poor but not classified as a fail the data were included in the final analysis but for an analysis aiming to quantify group differences, including this data would mean that a larger sample size would be required to detect true effects (Iscan et al. 2015). The process of performing visual QC and rejecting or editing poor quality data is likely to be easier than recruiting and scanning more patients. When investigating GM volume in a cohort with subtle disease effects such as pre-HD HD patients, the benefits of visual QC are likely to be substantial.

In addition to visual QC, manual editing can be performed where appropriate to improve results. In large cohorts this may be unfeasible, however as all techniques showed some segmentation errors in the TRACK-HD cohort and all offer options for manual editing this should be considered. It is particularly important for studies in which subtle group differences or longitudinal change is expected or when sub-regions are being examined. Manual editing requires the user to have in-depth knowledge of anatomy and to use a consistent procedure for specifying when manual editing should be performed.

One issue associated with visual QC is performing QC on ‘big data’ cohorts. Big data cohorts are increasing in availability, and consist of datasets whereby thousands of scans have been collected, for example the UK biobank study which aims to collect data from 100,000 participants (Alfaro-Almagro et al. 2018). This data is then often available to download and use. However, when processing data from so many participants visual QC is impossible. It is unclear how this issue can be managed, however since large datasets are likely to be more

robust to the effect of segmentation errors on analysis results, the issue may also be less important in these cohorts. While automated tools may be developed to assist with QC, it is unlikely that these will ever be as successful as visual QC performed by a trained human rater. UK Biobank is currently developing tools to perform automated QC (Alfaro-Almagro et al. 2018) and as these tools are validated against manual QC this issue will no doubt gain further attention.

3.4.1.6. How similar are results across the different tools?

For control participants all tools appeared to produce very similar results for both total GM and CGM; with between–technique correlation coefficients generally high, although slightly higher for CGM volume, indicating that techniques show greater variation in subcortical segmentation.

However, in HD participants the correlations were more variable, indicating that some techniques appear to detect disease-related change to a greater extent than others. In total GM volume, the relationships between techniques were generally lower than in controls. For CGM correlation coefficients were higher and more stable than for total GM, indicating that measurement of the subcortical grey matter may be more variable than CGM across different tools. As marked subcortical atrophy in the caudate and putamen is a defining feature of HD, it is unsurprising that the techniques may perform differently when segmenting this region, especially for tools developed on healthy controls. These results suggest that care should be taken when applying techniques in regions of severe atrophy or change, with much more divergent performance apparent in the use of the tools in these circumstances.

3.4.2. Limitations

The results of this study are limited by a few factors. Firstly, the aim was to examine segmentation procedures as non-technical users would routinely apply them and thus default brain extraction was used for each tool. This introduces additional variability to the comparison, and between-technique differences would likely be reduced if the same brain extraction was used prior to segmentation. In addition, the longitudinal pipelines offered by some tools were not compared (e.g. FreeSurfer, SPM, FSL) as not all tools currently offer a longitudinal pipeline.

There are a number of additional limitations in the examination of BrainWeb data. Firstly, the GT segmentations were not provided in the same resolution as the T1 scans that were downloaded; the T1 scans had a voxel size of 1x1x1mm whereas the GM volumes had a

voxel size of .5x.5x.5mm. Resampling was applied to the GM volumes to change the voxel size of the images. A number of different resampling methods were tested, with the most appropriate selected after careful examination of the output, and comparison of extracted GT volumes to previously published volumes for the GT segmentations showing an exact match (Klauschen et al. 2009). Finally, this comparison is limited by the absence of FreeSurfer in the partial volume analysis. FreeSurfer does not provide partial volume maps and so was excluded from the second analysis. Finally, in this study scans were used from multiple sites that could impact the performance of different techniques. However, this is also a relevant point given the increasing number of multi-site studies.

While as much consistency was maintained as possible during this analysis, some differences in performance were introduced by the differences described above, especially in the case of FreeSurfer. The developers of FreeSurfer recommend using the automatically output volumes provided in a text output for all participants. These values are calculated via correcting the volumetric regions for partial volume, and cannot easily be reproduced manually. Due to the difficulty in extracting GM volume from volumetric masks for FreeSurfer, total GM and CGM values were the values supplied automatically from FreeSurfer processing. Since the CGM volumes for FreeSurfer were output without overlaying the same mask used to extract values from the output of all other tools, the results cannot be directly compared to the other methods, and thus FreeSurfer is likely at an advantage for these results. For the lobular analysis, BrainWeb lobes were calculated using the automatic FreeSurfer methodology, but because the results were very different to the GT volumes, the TRACK-HD analysis used the lobular masks that were applied to the other segmentations. These were registered to FreeSurfer space and values were extracted using FreeSurfer commands. While this is possible, it required a number of posts on a FreeSurfer help list to output these values and the process was not always clear. All other techniques were examined using the partial volume maps in native space, with the same masks applied to the output of each technique.

3.4.3. Segmentation selection for the current work

When selecting a technique to use it is of course important that the question being asked is considered in relation to the type of measure provided by a tool; for the present research, CGM is the focus and the ideal measure would only output CGM without the need to use a manually registered mask to extract the cortical region. The use of a manual mask introduces a degree of measurement error as it is unlikely that manual masking of subcortical or cerebellar GM would be superior to a pre-optimised technique that extracts CGM. Another

issue to consider is that some measures do not provide regions that can be masked or that volumes can be extracted from, and indeed FreeSurfer makes this very difficult to do. It is not always clear how the volumes automatically output by the FreeSurfer pipeline are calculated, and it is often impossible to manually reproduce these volumes from the segmentations. If a proposed analysis does not use an already available FreeSurfer region then different software should be considered.

A final point to reiterate is regarding the use of these tools is how user-friendly they are. While most tools were fairly straightforward to apply, ANTs took a lot of time to understand and apply. Despite the large investment required to apply ANTs, the results were substandard, although it is possible that with a greater ability to optimise this tool the results could be improved. This tool may not always be appropriate for a novel user. In contrast, MALP-EM was the easiest method to use; it comes in an easy-to-install package, and since it is aimed at novice users there are no optional flags to adjust other than the use of a 1.5T vs. 3T flag.

Based on the qualitative and quantitative results MALP-EM demonstrated the most accurate and reliable performance for both the BrainWeb and TRACK-HD cohorts. MALP-EM will therefore be used for the CGM analysis in this thesis. It is hoped that because MALP-EM performed well on both datasets, is easy to use and freely available other neuroimaging users may be able to apply MALP-EM and undertake replication of the work completed in this thesis.

3.4.4. Conclusions

With a multitude of tools available to measure the volume of GM using MRI scans, the selection of the most appropriate tool can be a challenging first step in a research project, and one that may have a marked effect on the results of research. By using seven segmentation tools to quantify GM volume in a cohort including controls and participants with pathology, with both back-to-back and longitudinal data, a number of key points related to the measurement of GM volume could be highlighted. Table 10 provides some characteristics of the tools included in this study that can help to guide the initial decision making process in GM volumetric selection.

All methods compared in this work showed high reliability, supporting the results of previous studies and adding new information for the use of ANTs and MALP-EM, which have not been compared in this way previously. For the current cohort, the tools that detected the greatest raw change over a three-year follow-up period were not always the most sensitive to significant change in this period, indicating that higher variability in performance of these tools

reduced their sensitivity to subtle disease-related change. Despite this, all tools detected significant longitudinal change in GM when comparing the most advanced patients in the cohort to controls, meaning that all tools are sensitive to more advanced neural pathology across the whole cortex. Subtle regional differences were not detected by all tools, however. One of the most significant findings from this study is the importance of visual QC of GM segmentations. Poorly segmented and inaccurate data can easily be included in neuroimaging research if visual QC is omitted. While time consuming, the results of research that uses visual QC will be more sensitive and accurate than if it is not performed.

No one tool is more appropriate for analysis of every type of dataset or cohort. Several key considerations for the selection of the most appropriate grey matter segmentation tool have been identified. The scientific question, level of expertise and available resources are naturally paramount, while comparison of two or three different tools for ease and success of application is recommended and visual QC essential. For the current work, MALP-EM has been selected as the most appropriate GM segmentation tool.

4. EVALUATION OF CORTICAL THICKNESS MEASURES

In addition to volume, thickness of the CGM can be measured via MRI. CT is the distance between the WM/GM boundary and the GM/CSF boundary, and post mortem studies indicate that thickness ranges from approximately 1mm to 4.5mm, with an average across the cortex of 2.5mm (Fischl & Dale 2000). CT varies between individuals, shows changes across the lifespan and is hereditable (Panizzon et al. 2009; Fjell et al. 2015; Winkler et al. 2010; Winkler et al. 2009; Thambisetty et al. 2010; Storsve et al. 2014).

GM volume is a product of both the thickness and surface area of the cortex (another hereditable cortical characteristic genetically independent to CT), and thus it has been suggested that measuring cortical volume alone may not be sensitive enough to detect subtle disease-related brain changes, and that CT should also be examined (Panizzon et al. 2009). However, CT is challenging to calculate from MRI scans. Highly accurate delineations of the cortical surfaces are required, which is difficult to achieve in the complex sulci and gyri of brains. CT is calculated at every point in the cortex and errors in the delineation of the cortical surfaces can result in large overestimations of thickness. Even if overestimations only occur in a small number of regions they can result in large artificial inflations of global CT measures. These potential issues mean that the measurement tools used to calculate thickness are highly important.

CT is typically measured via either surface-based or voxel-based techniques, with an example of these two measures shown in Figure 4.1 (Clarkson et al. 2011). Surface-based measurement involves the construction of a meshed surface, typically on the GM/WM boundary, which is then expanded to encompass the outer surface between GM and CSF. The distance between these two surfaces can then be measured at each vertex or averaged across the whole cortex and sub-regions (Dale et al. 1999; Fischl et al. 1999; Fischl & Dale 2000). Surface-based methods tend to rely heavily on prior information about the brain to guide segmentation, known as probabilistic segmentation (Clarkson et al. 2011). Voxel-based techniques use the voxel-wise grid to first classify GM, WM and CSF and then measure the distance between WM and GM using Laplacian, registration, morphological or other techniques (Clarkson et al. 2011). Surface-based techniques are more computationally demanding, however they can overcome issues associated with the partial volume of voxels as the surface is not limited to follow voxel boundaries (Winkler et al. 2010). Due to this, they are considered more accurate than voxel-wise measures of thickness (Clarkson et al. 2011). Although they are affected by intensity non-uniformity and often perform poorly in areas with

low resolution and/or high curvature, due to the large amount of natural variation not covered in the probabilistic model. Furthermore, the use of probabilistic methods, which were created using a healthy-brain atlas, may not be appropriate for brains with pathological or anatomical differences. Despite these issues, surface-based methods are more often used to measure thickness largely due to a number of freely available and accessible measurement tools, with the most commonly used being FreeSurfer (Fischl & Dale 2000). Recently, however, voxel-based methods are becoming more popular due to the speed at which the data can be processed, lower segmentation failures and higher reliability than FreeSurfer (Schwarz et al. 2016).

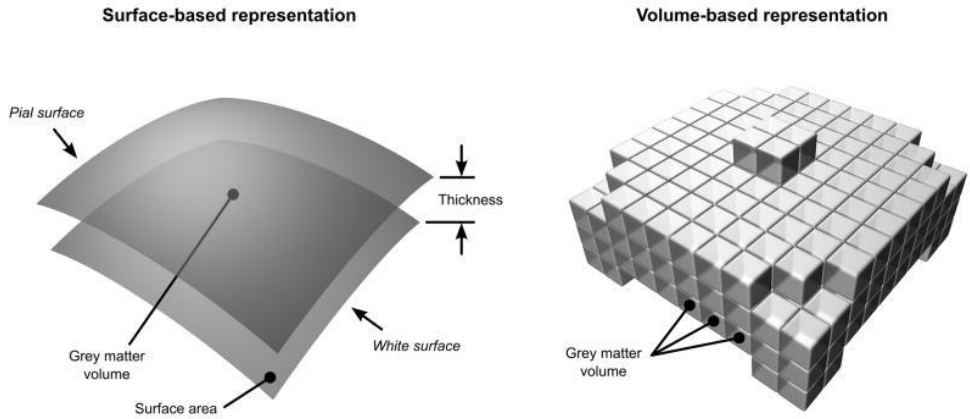


Figure 4.1. An example demonstrating the difference between surface- and volume-based methods of measuring CT (Winkler et al. 2010).

One major problem relevant to measuring CT is that of the failed segmentation of deep sulci (Clarkson et al. 2011). Deep, narrow sulci can be incorrectly classified as GM, resulting in an overestimation of thickness, as seen in Figure 4.2. If sulci are not separated via the segmentation process, CT is calculated spanning the two regions of CGM on either side of the sulci, creating in extremely large values for thickness at these points. This problem arises due to the inherently complicated pattern of cortical folding present in the brain, and although some methods aim to combat this problem via the use of limits that do not allow CT to be above a maximum biologically plausible threshold, it is a recognised difficulty within the field (Das et al. 2009; Clarkson et al. 2011).

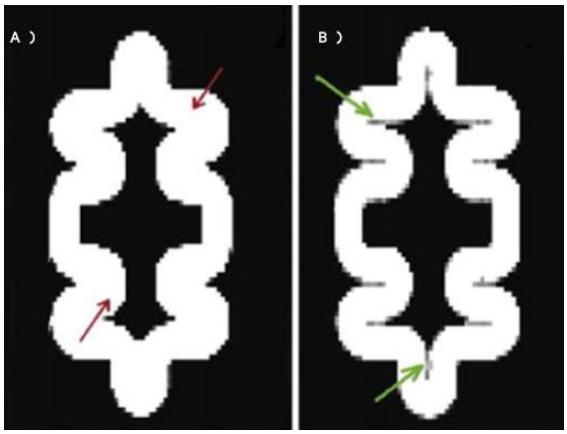


Figure 4.2. Examples of a segmentation whereby (A) sulci have been incorrectly classified as GM, shown by the red arrows, resulting in an overestimation of cortical thickness and (B) the sulci have been correctly classified, shown by the green arrows, resulting in a more accurate measure of cortical thickness (Clarkson et al. 2011).

Despite the use of MRI phantoms and models in a number of CT validation studies to try and assess accuracy of segmentation (Clarkson et al. 2011; Lee et al. 2006), it is difficult to establish a validate the performance of CT tools on real brain data. The complex nature of the convoluted cortex, the range of different definitions of the GM/CSF border and the WM/GM border, and different techniques and equations for calculating the distance between the two means that it is a complex metric to measure. Furthermore, post-mortem analyses of CT may not be accurate due to tissue shrinkage and so is not generally used as a comparison. Instead, thickness tools are often assessed for reliability, reproducibility and sensitivity. Some methods apply constraints in order to limit the maximal and minimal thickness values to values considered biologically plausible in order to try and aid accuracy, although using the same maximal and minimal values in healthy and very atrophied brains may not be appropriate (Das aet al. 2009).

A number of studies comparing CT measures have been performed, with some examining test-retest reliability of different tools (Iscan et al. 2015; Han et al. 2006; Wonderlick et al. 2009; Gronenschild et al. 2012; Schwarz et al. 2016) and some comparing the performance of thickness measures on their ability to separate participants by disease status (Tustison et al. 2014). FreeSurfer generally performs well when test-retest reliability is examined (Han et al. 2006; Jovicich et al. 2013; Liem et al. 2015), with previous studies reporting ICC values of >0.95 (Wonderlick et al. 2009). However, reliability for FreeSurfer varies across the cortex, with the orbito-temporal lobes often showing much lower reliability than other cortical regions (Han et al. 2006; Wonderlick et al. 2009; Schnack et al. 2010). A more recently developed voxel-based CT pipeline that uses ANTs software has given variable results during validation. When compared to FreeSurfer by the developers of ANTs, both

approaches showed good repeatability; with high intraclass coefficient values (ICC; ANTs=.98, FreeSurfer=.97) and low mean repeatability error for two databases (ANTs=3.2%, 3.3%, FreeSurfer=2.5%, 2.8%), with no significant difference between the two techniques for repeatability error. As a proxy for accuracy, the authors modelled age vs. CT using a 'training and test' paradigm to determine which measure produced better predictive ability. The ANTs model showed lower relative mean square error rates (Tustison et al. 2014). However subsequent validation has demonstrated that ANTs does not perform as well as initially reported, with reliability proving to be variable across different datasets (Tustison et al. 2014; Schwarz et al. 2016).

As with cortical volume, visual QC of CT measures is essential to detect failures in segmentation. Iscan et al. (2015) performed detailed QC on FreeSurfer segmentations for 40 participants conducted at two time points (80 total). They concluded that for 15 out of 40 participants there were errors in segmentation in either one time point or both that were severe enough to impact on the results. In contrast, Schwarz et al. (2016) reported that out of 452 scans processed with FreeSurfer, only 3 failed QC – although their criteria for failure was not provided. The authors also reported that FreeSurfer errors were seen more often in brains showing Alzheimer's pathology, suggesting a potential bias in thickness measurement. Few other studies have reported information on the quality of segmentations; however as with the measurement of GM volume it is vital that the most accurate tools are used to examine the thickness of the cortex in HD and other diseases. The process of performing QC on over 1000 FreeSurfer CT regions processed for the TRACK-HD and TrackOn-HD studies was performed prior to undertaking this thesis, and based on this experience it was clear that FreeSurfer CT regions are often inaccurate and output poorly delineated cortical surfaces. While CT offers a promising measure of cortical change for HD, the performance of CT tools on the TRACK-HD cohort required further investigation.

Despite these limitations, studies measuring CT have reported significant cortical thinning in pre-HD and HD participants (Rosas et al. 2002; Rosas et al. 2008; Harrington et al. 2014). When HD participants were separated into three groups based on their disease stage (with lower stages representing lower disease progression), the results suggested that thinning began in the primary motor area, sensory area, visual area and portions of the precuneus in stage I (Rosas et al. 2008). This thinning extended throughout the cortex in stage II and by stages III/IV most of the cortex had begun thinning, excluding the frontal cortex, which was relatively preserved. These studies are widely cited within the HD literature, however a number of the studies were conducted in small groups of participants, for example one study

had groups of N=8, N=14 and N=9 (Rosas et al. 2008). Furthermore, a number of widely cited results were not corrected for multiple comparisons (Rosas et al. 2008; Rosas et al. 2002). In addition, a recent study compared imaging biomarkers in the PADDINGTON cohort found that FreeSurfer was not a sensitive marker for detecting cortical change in manifest HD participants (Hobbs et al. 2015), raising doubt about the validity of using CT measures in HD studies.

4.1. Aims

This study aims to compare the performance of three CT measures on a subset of the TRACK-HD data. CT is a measure that potentially offers more sensitivity than volume; however it is a difficult characteristic to measure accurately. Prior to applying CT tools on the TRACK-HD data for examination of CGM they require detailed validation in this cohort.

4.2. Methods

4.2.1. Participants

The participants used in this study were the same participants used in the comparison of volumetric measures. 100 participants with data from 2008 and 2011 were randomly selected from the TRACK-HD study. Twenty control participants, 20 PreHD-A, 20 PreHD-B, 20 HD1 and 20 HD2 participants were included, see section 3.2.2.2 for further information, and Table 3.5 for demographics information.

4.2.2. Cortical thickness measurements

Three methods of measuring CT were examined: A) CT measured via FreeSurfer, B) CT measured via ANTs, C) CT measured using MALP-EM volumetric segmentations input to the ANTs Kelly Kapowski pipeline that calculates CT from volumetric regions. These methods were selected due to a number of considerations. FreeSurfer is the most commonly used software for CT measurements, despite concerns about the accuracy of FreeSurfer segmentations, and was thus included in this comparison. As previously mentioned, ANTs has been more recently developed yet it has been used in a growing number of publications characterising CT. It has a pipeline that includes brain extraction, segmentation and measurement of thickness. Finally, the cortical segmentations from Chapter 3 delineated using MALP-EM were combined with the calculation of CT implemented in ANTs.

The longitudinal analysis was conducted twice: first, the two longitudinal scans were analysed in native space. This resulted in large longitudinal change in CT measures for some tools, with biologically implausible changes in thickness found. Since most automatic

longitudinal pipelines either include or recommend a registration step, the data was also registered and re-analysed. The 2008 and 2011 scans were registered using the SPM longitudinal registration pipeline described in 2.3.1.2.3 (Ashburner & Ridgway 2012). Default parameters were used to create an average image for each participant. The warping parameters were applied to the native scans to align them with the template for each participant, and these warped scans were then segmented independently.

FreeSurfer version 5.3.0 was used, with the default processing pipeline. MALP-EM was run using the default settings, with a flag to specify 3T scans. Following this, the ANTs KellyKapowski pipeline was run on the MALP-EM regions using default settings. ANTs was run using the study-specific templates created in section 3.2.3.4.2, with the CT command, antsCorticalThickness.sh, then applied with default settings.

Following registration and any processing, images were visually checked for segmentation errors, with only gross-errors as described in section 3.2.3 excluding scans from analysis.

4.3. Results

4.3.1. Quality Control

As in Chapter 3, visual QC was important for interpretation of the thickness results. Upon inspection, ANTs segmentations were inconsistent and a high number of segmentations had considerable errors, especially within the occipital and temporal GM. The most common errors were the over-inclusion of non-brain tissue in the cortical segmentation and were particularly pronounced for brains showing a lot of atrophy, as shown in Figure 4.3. The quality of MALP-EM segmentations was better than ANTs, with minor over-inclusion occasionally seen in occipital and temporal lobes (Figure 4.4). FreeSurfer regions were also variable, with errors in both the WM boundary and the external CSF boundary common (Figure 4.5). The FreeSurfer segmentations had less obvious errors in segmentation, yet there were often large chunks excluded from both internal GM (WM overestimation extended into the GM) and more external GM (WM classified as CSF). In addition, the temporal lobes were often poorly delineated as was the occipital lobe.

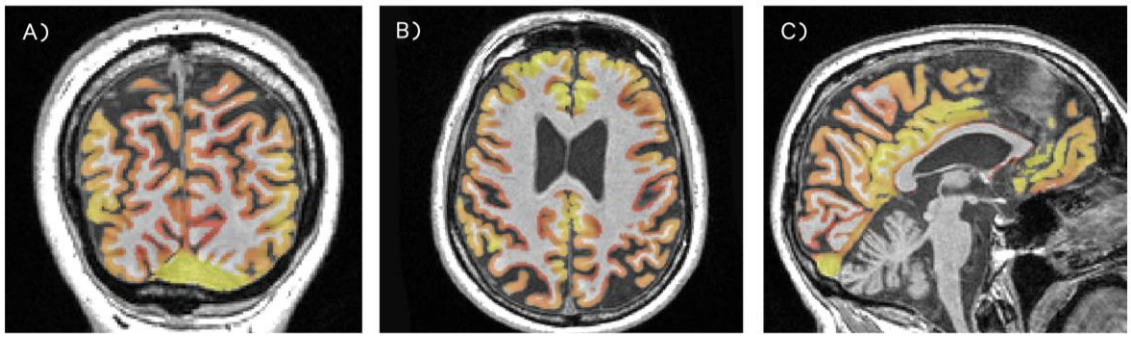


Figure 4.3. An example of an ANTs CT segmentation for a participant showing disease related atrophy shown in (A) coronal, (B) axial and (C) sagittal views. Spillage of the CT segmentation into the occipital dura is particularly pronounced in the coronal and sagittal views.

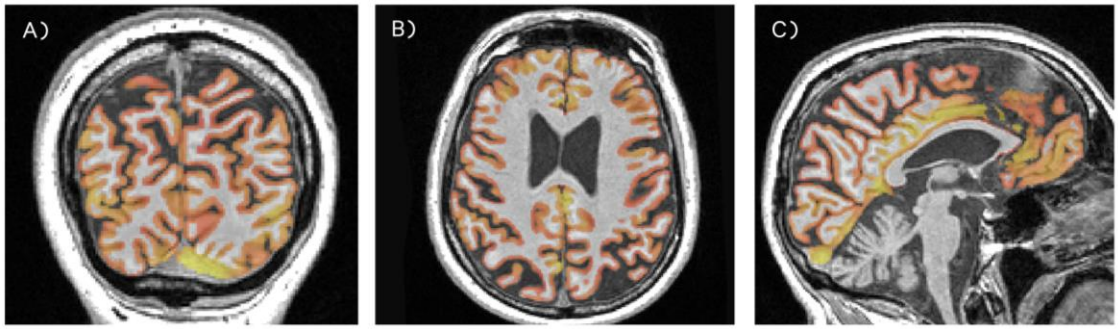


Figure 4.4. An example of a MALP-EM CT segmentation for a participant showing disease related atrophy shown in A) coronal, (B) axial and (C) sagittal views. Spillage of the CT segmentation into the occipital dura is particularly pronounced in the coronal and sagittal views.

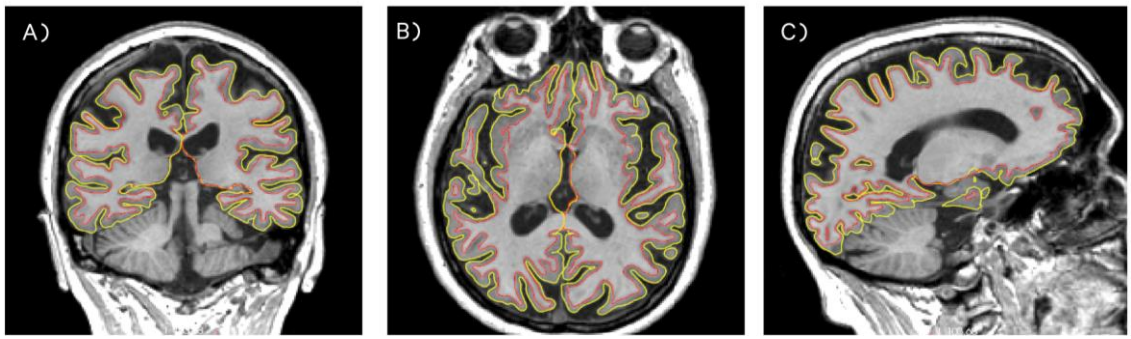


Figure 4.5. An example of a FreeSurfer CT segmentation for a participant showing disease related atrophy shown in A) coronal, (B) axial and (C) sagittal views. Underestimation of the GM is particularly pronounced in the coronal and sagittal views.

In addition to these issues, a more widely accepted problem previously recognised in CT measurement was seen during QC. That is, it is very difficult for most tools to accurately define the most tightly bound sulci. This is often more problematic in voxel-based techniques, and examples from the current study can be seen in Figure 4.6, compared to the processing performed via FreeSurfer, the only surface-based technique included in this comparison.

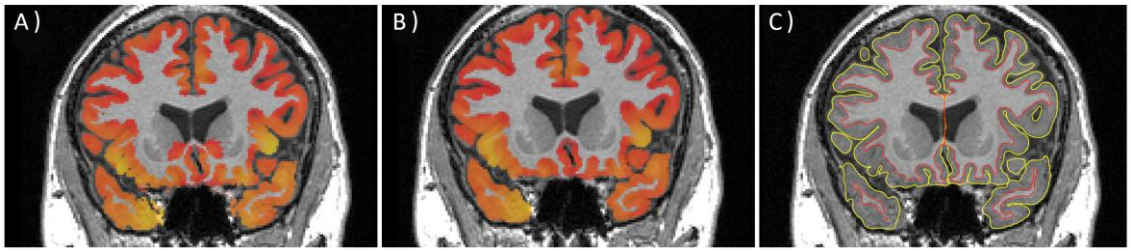


Figure 4.6. An example of (A) an ANTs cortical thickness segmentation, (B) a MALP-EM cortical segmentation processed with the ANTs cortical thickness calculation, and (C) a FreeSurfer segmentation. In both the ANTs and MALP-EM segmentations, sulci are not segmented accurately resulting in overestimation of thickness. This can be seen clearly in the inferior regions of this scan.

4.3.2. Cross-sectional results

4.3.2.1. Total thickness

Table 4.1 shows the results of the test-retest analysis for the three processing pipelines, separated into HD groups and controls. For all tools, CT was lower in HD1 and HD2 than for pre-HD and control participants, but there was inconsistency within both control and pre-HD groups across different tools. Overall, FreeSurfer had lower thickness values than both ANTs and MALP-EM for all groups.

The ICC for ANTs was lower than for both MALP-EM and FreeSurfer, ranging from 0.70-0.87, and repeatability was above 5.00 for all groups. However, confidence intervals (CIs) for the ICC ranged from around 0 to >.90, indicating poor reliability. MALP-EM showed better reliability, with ICC values above 0.96 (and CIs >.90) and repeatability less than 1.5 for all groups. Finally, FreeSurfer ICC values ranged from 0.819-0.944, but with one lower CI falling to .599, indicating that true reliability may be lower. Repeatability values were between 0.96-1.67.

Table 4.1. Mean, standard deviations and ranges for cortical thickness in both 2008 scans for each group and technique. Intraclass correlations and confidence intervals between these scans are also presented, as well as repeatability means, standard deviations and ranges.

		Scan A	Scan B	ICC	Repeatability
ANTs	**Controls**	2.96 (0.26)	2.77 (0.30)	0.751	7.12 (3.63)
		2.44-3.47	2.23-3.33	(-0.067-0.936)	2.04-13.81
	PreHD-A	2.94 (0.25)	2.76 (0.30)	0.775	6.77 (3.76)
		2.40-3.35	2.09-3.20	(-0.062-0.946)	0.66-14.08
	PreHD-B	2.88 (0.21)	2.71 (0.21)	0.700	6.23 (2.78)
		2.50-3.30	2.32-3.14	(-0.068-0.923)	1.80-11.61
	HD1	2.73 (0.30)	2.58 (0.30)	0.870	5.90 (2.35)
		2.42-3.46	2.24-3.30	(-0.031-0.973)	2.38-11.93
	HD2	2.75 (0.28)	2.55 (0.30)	0.776	7.44 (3.62)
		2.36-3.40	2.11-3.30	(-0.061-0.946)	2.83-15.51
MALP-EM + ANTs	**Controls**	2.89 (0.27)	2.90 (0.27)	0.992	0.93 (0.81)
		2.42-3.49	2.37-3.49	(0.980-0.997)	0.03-2.59
	PreHD-A	2.93 (0.19)	2.91 (0.20)	0.978	1.03 (0.98)
		2.56-3.22	2.52-3.25	(0.945-0.991)	0.01-3.88
	PreHD-B	2.96 (0.15)	2.96 (0.17)	0.962	1.05 (1.13)
		2.73-3.29	2.70-3.38	(0.906-0.985)	0.01-4.14
	HD1	2.84 (0.29)	2.84 (0.27)	0.985	1.28 (1.61)
		1.99-3.30	2.14-3.29	(0.962-0.994)	0.00-7.16
	HD2	2.79 (0.27)	2.78 (0.28)	0.981	1.34 (1.87)
		2.03-3.20	1.89-3.19	(0.952-0.993)	0.02-7.11
FreeSurfer	**Controls**	2.46 (0.10)	2.45 (0.09)	0.836	1.67 (1.61)
		2.31-2.66	2.29-2.65	(0.637-0.931)	0.07-5.91
	PreHD-A	2.42 (0.08)	2.41 (0.09)	0.934	1.01 (0.75)
		2.31-2.58	2.29-2.57	(0.842-0.973)	0.01-2.30
	PreHD-B	2.43 (0.08)	2.41 (0.08)	0.819	1.46 (1.39)
		2.30-2.63	2.30-2.55	(0.599-0.924)	0.03-4.75
	HD1	2.35 (0.10)	2.34 (0.09)	0.927	1.13 (1.01)
		2.20-2.56	2.19-2.54	(0.800-0.972)	0.21-4.80
	HD2	2.33 (0.09)	2.32 (0.09)	0.944	0.96 (0.94)
		2.15-2.46	2.10-2.47	(0.864-0.978)	0.13-3.28

4.3.2.2. Regional Thickness

Table 4.2 shows lobular values for 2008 Scan A and Scan B, and Table 4.3 shows ICC and repeatability for these scans. For ANTs CT, the lobular repeatability metrics were generally poor. Average thickness values varied, and ICC scores ranged from .456 in the insula (control participants) to .927 in the parietal lobe (HD1 participants). Confidence intervals for ICC values were very broad, again indicating that reliability was inadequate when using this technique in this cohort. Repeatability was low, with the highest values (representing poor repeatability) seen in the insula and occipital lobe.

For MALP-EM, ICC values were again high, with all >.95, and no pattern of lower ICC in a particular region. In addition to the high ICC values, the confidence intervals for all regions were also very high. Repeatability values were lower than 2 for all regions except for the temporal lobe in HD2, which had a value of 2.57. For FreeSurfer, ICC values ranged from around .71 to .96, with no clear lobular pattern. However, CIs for the ICC measures showed a

wide range, indicating possible lower reliability. Repeatability values were generally <2, with the highest value seen in the insula for HD2 (2.93).

4.3.3. Longitudinal results

4.3.3.1. Non-registered

The results from the analysis examining non-registered longitudinal data provided in Table 4.4 and Figure 4.7 demonstrated a number of key findings. Firstly, ANTs showed a large amount of longitudinal change within controls undergoing a reduction in CT of 3.42%. This was larger than the change seen in both PreHD-A and PreHD-B, however HD1 and HD2 both underwent greater reductions in CT (5.07% and 4.75% respectively). Only HD1 showed significantly greater longitudinal change than controls, at p=.044. The range of CT change for each group was large, however, with volume reduction of up to 15.62% seen in one HD1 participant, and thickness increases of 5.66% seen in one PreHD-B participant. This result reflects the poor reliability seen in ANTs in the cross-sectional analyses.

For MALP-EM, there were extremely wide ranges in the amount of change occurring over the 3 year period; with changes of up to 42% seen in HD1 (both increases and decreases in thickness). While controls showed a low mean rate of change (0.15%), the range for this group was also substantial (19% increases and 18% decreases could be seen). No group showed significantly different change when compared to controls, although all groups underwent a higher rate of change.

The range of FreeSurfer rates of change were narrower, with controls showing very low change and HD groups undergoing progressively greater thinning of the cortex with increasing disease burden. PreHD-B, HD1 and HD2 all underwent significantly greater CT reductions than controls, with HD2 undergoing the greatest change. The ranges for FreeSurfer were more in line with expectations, showing generally small to moderate change in thickness over three years.

4.3.3.1. Registered

The scans were registered and re-analysed using the same pipeline. Results are shown in Figure 4.5 and Figure 4.7. After registering the scans and re-analysing the data using ANTs, controls underwent a large amount of change showing a greater reduction in thickness than any HD group (3.60%). There was no significant difference between controls and any HD group, with no pattern of greater thickness reduction as disease burden increased.

The results from MALP-EM indicated that registration did not improve longitudinal performance. Controls showed a large decrease in volume over 3 years (2.94%), with PreHD-A only undergoing .25% reduction. PreHD-B, HD1 and HD2 all underwent reductions >3.5%. For most groups, the ranges were reduced compared to non-registered scans, although still wide.

After running FreeSurfer on the registered T1 scans, there was no longer a pattern of increasing atrophy with increasing disease burden. Again, controls showed greater change than any other group (1.54%), and similarly to other methods there was no pattern of increasing rate of change with increasing disease burden.

Table 4.2. Mean, standard deviations and ranges of cortical thickness values across different lobes for both 2008 scans for each technique.

		Controls		PreHD-A		PreHD-B		HD1		HD2	
		Scan A	Scan B	Scan A	Scan B	Scan A	Scan B	Scan A	Scan B	Scan A	Scan B
ANTs	Frontal	2.73 (0.28) 2.32-3.21	2.55 (0.33) 2.12-3.13	2.71 (0.27) 2.17-3.16	2.53 (0.32) 1.82-3.10	2.69 (0.21) 2.35-3.11	2.55 (0.20) 2.20-2.87	2.59 (0.29) 2.24-3.27	2.44 (0.28) 2.06-3.10	2.65 (0.29) 2.24-3.16	2.50 (0.31) 2.02-3.07
	Occipital	2.86 (0.28) 2.19-3.56	2.64 (0.32) 1.97-3.42	2.78 (0.36) 2.14-3.26	2.57 (0.35) 1.90-3.09	2.68 (0.35) 2.26-3.57	2.45 (0.33) 2.02-3.43	2.38 (0.41) 1.89-3.31	2.20 (0.42) 1.69-3.19	2.27 (0.39) 1.83-3.39	2.04 (0.39) 1.64-3.27
	Parietal	2.51 (0.24) 2.06-2.91	2.37 (0.25) 1.86-2.80	2.48 (0.24) 1.98-2.88	2.35 (0.25) 1.73-2.76	2.41 (0.17) 2.14-2.72	2.29 (0.16) 2.02-2.61	2.26 (0.27) 1.93-2.91	2.17 (0.27) 1.83-2.83	2.31 (0.23) 1.98-2.76	2.18 (0.23) 1.80-2.70
	Temporal	3.72 (0.29) 3.10-4.26	3.48 (0.34) 2.83-4.04	3.70 (0.30) 3.12-4.24	3.47 (0.35) 2.75-4.06	3.64 (0.30) 3.15-4.29	3.42 (0.29) 2.95-4.00	3.48 (0.34) 3.07-4.34	3.28 (0.37) 2.84-4.13	3.47 (0.35) 3.00-4.22	3.19 (0.38) 2.67-4.10
	Insula	3.53 (0.33) 2.99-4.09	3.10 (0.34) 2.51-3.00	3.74 (0.41) 3.02-4.59	3.35 (0.42) 2.72-4.25	3.60 (0.35) 2.74-4.25	3.22 (0.36) 2.49-3.81	3.46 (0.28) 2.98-3.99	3.15 (0.37) 2.56-3.87	3.62 (0.45) 3.00-4.43	3.21 (0.47) 2.59-4.25
MALP-EM + ANTs	Frontal	2.71 (0.27) 2.19-3.25	2.73 (0.26) 2.20-3.23	2.74 (0.20) 2.29-3.07	2.74 (0.22) 2.26-3.14	2.78 (0.17) 2.55-3.19	2.78 (0.18) 2.49-3.14	2.67 (0.23) 1.99-3.05	2.68 (0.22) 2.13-3.04	2.66 (0.26) 1.97-3.13	2.66 (0.26) 1.87-3.11
	Occipital	2.77 (0.29) 2.42-3.31	2.78 (0.30) 2.41-3.35	2.78 (0.28) 2.30-3.21	2.76 (0.28) 2.25-3.21	2.84 (0.35) 2.16-3.61	2.83 (0.36) 2.13-3.65	2.60 (0.38) 1.77-3.20	2.60 (0.37) 1.75-3.19	2.42 (0.32) 1.77-3.13	2.41 (0.32) 1.68-3.11
	Parietal	2.61 (0.26) 2.21-3.31	2.63 (0.26) 2.22-3.32	2.62 (0.19) 2.23-2.88	2.60 (0.19) 2.22-2.88	2.57 (0.14) 2.39-2.84	2.58 (0.16) 2.30-2.95	2.48 (0.25) 1.85-2.92	2.48 (0.24) 1.95-2.91	2.44 (0.21) 1.83-2.74	2.44 (0.22) 1.67-2.71
	Temporal	3.53 (0.32) 2.88-4.12	3.51 (0.32) 2.79-4.13	3.57 (0.25) 3.15-4.04	3.53 (0.26) 3.17-4.06	3.64 (0.22) 3.19-4.16	3.64 (0.24) 3.19-4.20	3.55 (0.45) 2.23-4.15	3.54 (0.40) 2.49-4.11	3.49 (0.38) 2.48-4.06	3.45 (0.41) 2.27-4.09
	Insula	3.38 (0.45) 2.52-4.18	3.38 (0.47) 2.51-4.23	3.69 (0.53) 2.84-4.57	3.68 (0.52) 2.90-4.63	3.74 (0.55) 3.04-5.16	3.76 (0.56) 3.00-5.23	3.50 (0.49) 2.19-4.36	3.50 (0.46) 2.31-4.22	3.52 (0.63) 2.30-4.73	3.50 (0.65) 2.10-4.69
FreeSurfer	Frontal	2.50 (0.11) 2.35-2.71	2.48 (0.10) 2.36-2.70	2.45 (0.09) 2.31-2.62	2.44 (0.10) 2.29-2.63	2.46 (0.09) 2.33-2.60	2.45 (0.08) 2.31-2.59	2.41 (0.09) 2.21-2.56	2.39 (0.09) 2.23-2.55	2.40 (0.10) 2.19-2.55	2.39 (0.12) 2.13-2.57
	Occipital	1.98 (0.10) 1.83-2.20	1.98 (0.10) 1.80-2.20	1.97 (0.11) 1.81-2.19	1.96 (0.11) 1.75-2.19	1.95 (0.10) 1.78-2.23	1.93 (0.10) 1.71-2.17	1.85 (0.11) 1.67-2.07	1.83 (0.11) 1.66-2.04	1.76 (0.09) 1.62-1.91	1.76 (0.10) 1.61-1.92
	Parietal	2.85 (0.13) 2.62-3.12	2.84 (0.12) 2.57-3.04	2.81 (0.11) 2.63-3.05	2.80 (0.11) 2.63-2.99	2.80 (0.10) 2.68-3.11	2.79 (0.11) 2.61-3.03	2.72 (0.14) 2.43-3.01	2.72 (0.13) 2.45-3.01	2.71 (0.10) 2.50-2.87	2.70 (0.11) 2.48-2.92
	Temporal	2.30 (0.11) 2.10-2.53	2.28 (0.10) 2.14-2.54	2.25 (0.08) 2.13-2.37	2.24 (0.09) 2.08-2.42	2.27 (0.10) 2.11-2.45	2.24 (0.09) 2.12-2.45	2.17 (0.11) 2.02-2.40	2.15 (0.10) 2.02-2.37	2.14 (0.10) 1.96-2.31	2.13 (0.10) 1.91-2.31
	Insula	2.98 (0.15) 2.78-3.33	2.99 (0.14) 2.73-3.28	3.03 (0.11) 2.86-3.28	3.03 (0.10) 2.92-3.24	3.01 (0.15) 2.77-3.29	2.99 (0.15) 2.71-3.30	3.01 (0.13) 2.69-3.24	3.01 (0.11) 2.79-3.16	2.96 (0.17) 2.65-3.22	2.93 (0.16) 2.65-3.19

Table 4.3. Intraclass correlations and confidence intervals, and repeatability means, standard deviations and ranges for both 2008 scans when measuring cortical thickness across different techniques and regions.

		Controls		PreHD-A		PreHD-B		HD1		HD2	
		ICC	**Repeatability**	**ICC**	**Repeatability**	**ICC**	**Repeatability**	**ICC**	**Repeatability**	**ICC**	**Repeatability**
ANTs	**Frontal**	0.806 (-0.052-0.953)	7.15 (3.93) 0.93-14.83	0.825 (-0.047-0.959)	6.77 (4.13) 0.35-17.49	0.764 (-0.052-0.945)	5.65 (2.18) 1.65-9.83	0.885 (-0.027-0.977)	5.64 (2.21) 1.69-9.09	0.839 (-0.032-0.962)	6.27 (3.72) 0.33-13.95
	Occipital	0.716 (-0.075-0.924)	8.47 (4.90) 0.72-19.19	0.796 (-0.058-0.951)	8.48 (4.37) 1.86-16.36	0.777 (-0.059-0.947)	9.04 (3.94) 1.14-14.66	0.893 (-0.024-0.979)	8.58 (3.70) 3.15-20.19	0.835 (-0.042-0.964)	10.71 (4.18) 3.54-18.21
	Parietal	0.789 (-0.029-0.944)	6.15 (3.70) 1.89-16.54	0.843 (-0.037-0.964)	5.63 (3.35) 0.26-13.37	0.723 (-0.073-0.928)	5.16 (2.56) 0.77-9.89	0.927 (0.006-0.985)	4.53 (2.04) 2.36-8.60	0.821 (-0.040-0.957)	5.81 (3.33) 0.69-13.38
	Temporal	0.707 (-0.076-0.922)	7.24 (3.28) 2.71-12.85	0.740 (-0.070-0.934)	6.83 (3.87) 0.69-14.07	0.720 (-0.074-0.927)	6.32 (3.25) 1.89-14.21	0.824 (-0.037-0.963)	6.34 (2.35) 3.33-12.07	0.724 (-0.066-0.931)	8.63 (4.09) 2.89-17.10
	Insula	0.456 (-0.080-0.811)	13.04 (6.30) 6.19-25.20	0.615 (-0.090-0.886)	10.90 (6.32) 0.43-20.49	0.522 (-0.094-0.842)	11.45 (6.32) 1.60-26.02	0.631 (-0.090-0.892)	9.48 (5.45) 2.34-22.47	0.632 (-0.086-0.895)	12.21 (6.48) 1.54-25.19
MALP-EM + ANTs	**Frontal**	0.984 (0.961-0.994)	1.32 (1.19) 0.00-3.61	0.968 (0.921-0.987)	1.26 (1.35) 0.09-5.61	0.946 (0.870-0.978)	1.49 (1.47) 0.01-6.46	0.968 (0.922-0.987)	1.79 (1.54) 0.19-6.84	0.983 (0.957-0.993)	1.44 (1.45) 0.15-5.31
	Occipital	0.987 (0.969-0.995)	1.34 (1.07) 0.29-4.39	0.977 (0.942-0.991)	1.77 (1.52) 0.02-5.65	0.989 (0.972-0.996)	1.39 (1.24) 0.01-5.54	0.991 (0.978-0.997)	1.47 (1.21) 0.19-3.87	0.992 (0.979-0.997)	1.49 (1.30) 0.13-5.61
	Parietal	0.988 (0.965-0.995)	1.13 (1.11) 0.03-3.91	0.972 (0.925-0.989)	1.21 (1.22) 0.09-5.07	0.939 (0.855-0.975)	1.43 (1.34) 0.04-5.36	0.987 (0.967-0.995)	1.23 (1.30) 0.00-5.40	0.969 (0.923-0.988)	1.61 (2.09) 0.14-8.77
	Temporal	0.986 (0.962-0.995)	1.13 (1.11) 0.03-3.62	0.966 (0.897-0.988)	1.25 (1.49) 0.00-5.13	0.965 (0.915-0.986)	1.27 (1.12) 0.03-4.75	0.984 (0.961-0.994)	1.59 (2.39) 0.12-11.04	0.963 (0.906-0.986)	2.57 (2.37) 0.09-8.54
	Insula	0.985 (0.963-0.994)	1.73 (1.39) 0.38-4.76	0.987 (0.968-0.995)	1.82 (1.47) 0.07-5.94	0.990 (0.974-0.996)	1.71 (1.22) 0.21-5.14	0.986 (0.966-0.995)	1.75 (1.64) 0.04-5.78	0.995 (0.986-0.998)	1.57 (1.99) 0.08-8.88
FreeSurfer	**Frontal**	0.710 (0.406-0.874)	2.06 (2.34) 0.01-8.50	0.932 (0.836-0.972)	1.34 (1.16) 0.01-3.69	0.914 (0.799-0.965)	1.91 (1.67) 0.08-5.97	0.796 (0.558-0.914)	1.38 (1.14) 0.10-5.02	0.916 (0.803-0.966)	1.01 (1.02) 0.11-3.51
	Occipital	0.893 (0.754-0.956)	1.61 (1.12) 0.08-4.48	0.966 (0.918-0.986)	1.23 (0.91) 0.00-3.56	0.911 (0.792-0.963)	1.65 (1.20) 0.30-4.56	0.924 (0.821-0.969)	1.27 (0.77) 0.15-2.61	0.784 (0.529-0.909)	0.90 (0.56) 0.21-2.38
	Parietal	0.721 (0.424-0.879)	1.29 (1.31) 0.02-5.24	0.922 (0.785-0.970)	1.18 (1.19) 0.05-4.52	0.858 (0.679-0.941)	1.51 (1.35) 0.00-4.10	0.807 (0.565-0.920)	0.95 (0.77) 0.08-3.56	0.851 (0.667-0.938)	1.16 (1.18) 0.01-4.09
	Temporal	0.893 (0.703-0.959)	2.13 (2.11) 0.18-7.03	0.969 (0.687-0.992)	1.08 (0.99) 0.06-3.24	0.969 (0.925-0.988)	2.05 (1.70) 0.14-7.03	0.878 (0.699-0.951)	1.55 (1.80) 0.17-7.73	0.827 (0.613-0.928)	1.18 (1.28) 0.08-3.78
	Insula	0.958 (0.894-0.983)	1.71 (1.02) 0.47-3.90	0.981 (0.953-0.993)	1.90 (1.10) 0.20-3.92	0.920 (0.809-0.968)	2.19 (1.74) 0.12-6.09	0.934 (0.840-0.974)	1.92 (1.45) 0.36-6.29	0.810 (0.578-0.921)	2.63 (2.29) 0.06-9.05

Table 4.4. Mean % change (difference between 2011 and baseline thickness as a percentage of baseline thickness), standard deviation and ranges for all tools and groups in average cortical thickness. Positive values represent thickness decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, with significantly greater change in HD groups represented by * p<0.05 (light grey), **p<0.01 (dark grey). Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
ANTs	**% Decrease 2008 to 2011**	3.42 (2.92) -0.65-9.41	3.20 (4.08) -3.89-14.31	2.52 (3.66) -5.66-9.70	5.07 (3.26) 2.31-15.62	4.75 (3.13) -3.59-10.95
	Significant difference		1.09 (-2.11-2.18) p=0.974	-0.48 (-1.36-0.40) p=0.284	0.56 (0.01-1.11) p=0.044*	0.36 (-0.08-0.79) P=0.109
MALP-EM + ANTs	**% Decrease 2008 to 2011**	0.15 (7.32) -19.72-18.05	2.94 (5.91) -2.65-22.59	2.67 (7.29) -6.03-25.94	3.89 (16.20) -42.25-42.24	2.49 (5.79) -6.92-22.58
	Significant difference		2.97 (-1.08-7.01) p=0.150	0.92 (-1.12-2.96) p=0.375	0.88 (-1.56-3.32) p=0.480	0.15 (-1.01-1.31) p=0.798
FreeSurfer	**% Decrease 2008 to 2011**	0.10 (1.12) -2.11-2.34	0.65 (1.22) -1.48-3.17	1.00 (1.57) -1.36-3.44	2.03 (1.62) -1.26-4.91	2.33 (1.81) -0.98-5.57
	Significant difference		0.37 (-0.31-1.05) p=0.289	0.45 (0.09-0.82) p=0.015*	0.54 (0.27-0.82) p<0.001**	0.46 (0.21-0.70) p<0.001**

Table 4.5. Mean % change in registered scan pairs (difference between 2011 and baseline thickness as a percentage of baseline thickness), standard deviation and ranges for all tools and groups in average cortical thickness. Positive values represent thickness decreases over time. Results of regression analyses comparing rate of change in controls to HD groups, no significant group differences were found. Age, sex and site were controlled for.

		Controls	PreHD-A	PreHD-B	HD1	HD2
ANTs	**% Decrease 2008 to 2011**	3.60 (3.10) -0.62-11.20	3.28 (4.71) -3.56-13.47	2.28 (3.20) -4.41-11.15	1.48 (4.47) -8.37-14.26	3.13 (2.15) 0.13-6.79
	Significant difference		.18 (-2.17-2.54) p=0.880	-.65 (-1.56-.26) p=0.163	-.62 (-1.37-.12) p=0.101	.05 (-.44-.55) P=0.829
MALP-EM + ANTs	**% Decrease 2008 to 2011**	2.94 (8.15) -25.24-19.08	0.25 (9.43) -22.13-13.96	3.58 (4.14) -5.97-9.86	4.28 (7.25) -15.15-19.42	3.98 (4.82) -4.19-14.37
	Significant difference		-1.68 (-7.08-3.72) p=.542	.52 (-1.23-2.27) p=.561	.69 (-.78-2.15) p=.358	.85 (-.30-2.00) p=.148
FreeSurfer	**% Decrease 2008 to 2011**	1.54 (2.05) -2.21-5.39	0.89 (1.89) -2.15-4.57	1.44 (1.76) -0.49-4.86	0.61 (1.53) -3.64-3.36	1.24 (2.12) -3.55-5.02
	Significant difference		-.29 (-1.56-.97) p=.648	.04 (-.55-.63) p=.893	-.28 (-.66-.10) p=.142	-.07 (-.46-.32) p=.733

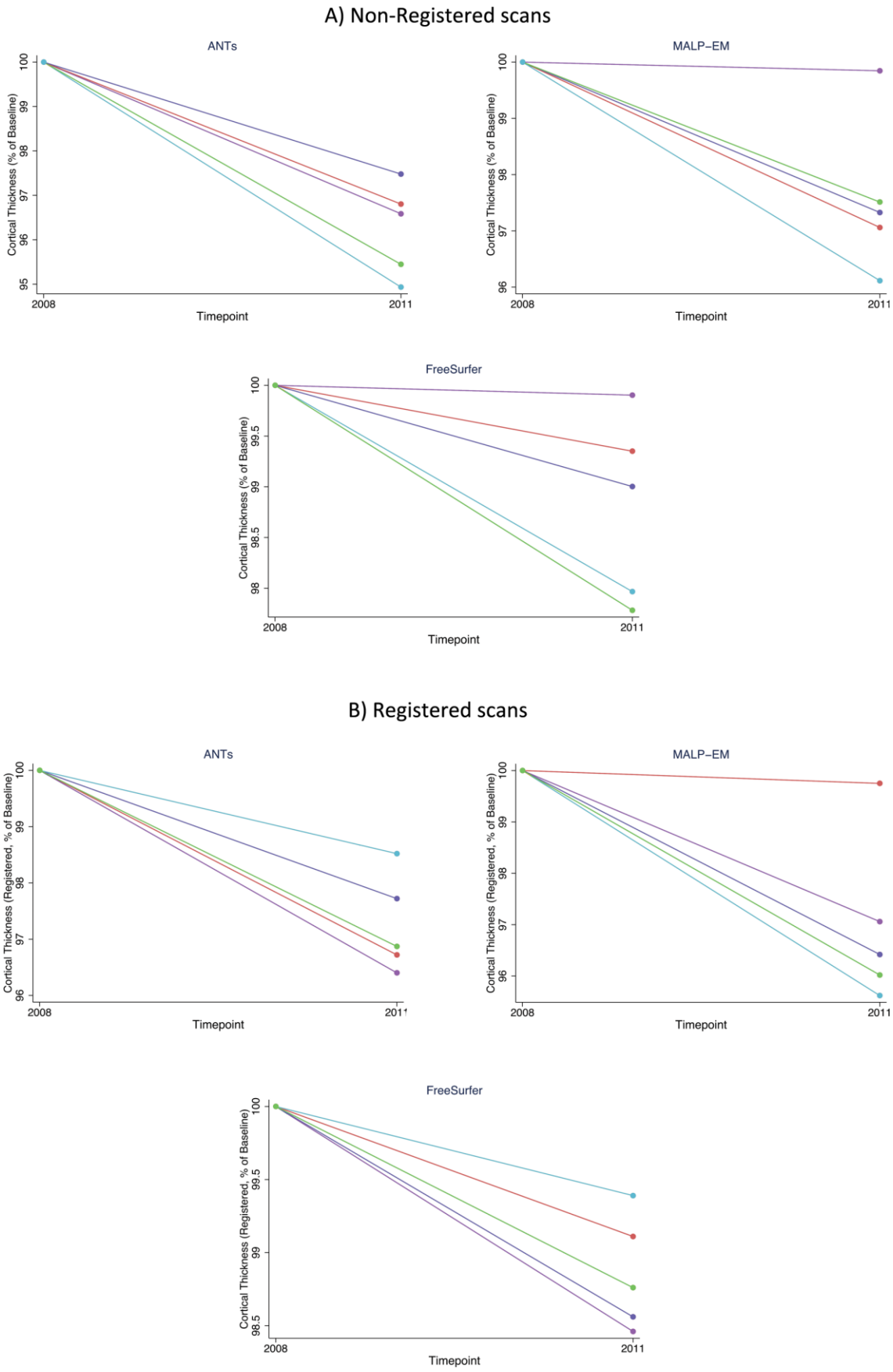


Figure 4.7. The longitudinal results showing change in cortical thickness from 2008 to 2011 time points for both A) non-registered and B) registered scan pairs.

4.4. Discussion

The results of this study have highlighted that there are a number of serious concerns regarding the current tools used to measure CT from MRI scans. Historically, there have been two ways to measure CT: surface-based techniques, such as FreeSurfer, and voxel-wise techniques, such as ANTs. While all measures have benefits and drawbacks, these results demonstrate that reliability and accuracy remain key issues with both general techniques.

FreeSurfer, the most commonly used CT measure, showed reasonable cross-sectional reliability and longitudinal sensitivity in non-registered scans. However, visual inspection of the segmentations indicated that the accuracy of tools was poor in a number of regions. The temporal lobes were generally poorly delineated, as has been previously reported (McCarthy et al. 2015). In addition, there was often incorrect classification of the cortical GM as WM or CSF. Despite these issues, when using the standard pipeline, FreeSurfer did detect significantly greater reduction in CT for gene carriers in PreHD-B, HD1 and HD2 compared to controls. This indicates some level of sensitivity to change, however it is impossible to tell from the current analysis how accurate these results are given the quality of many of the segmentations. There is particular concern when also considering the results of the longitudinal analysis measuring CT in registered scans, whereby the pattern of greater reductions in CT with increasing disease burden was not apparent and a number of participants showed very different rates of longitudinal change compared to those seen for the non-registered scans. After the statistical analysis was performed showing poor longitudinal results, a selection of segmentations were visually re-examined. There was no obvious reason for the differing performance of FreeSurfer after registration, that is, there were no visible errors in segmentation in registered scan pairs. The different longitudinal results seen in non-registered and registered scans extended across all lobes, indicating that these differences were not due to errors in one region. These results suggest that by performing registration on the scans prior to segmentation the act of warping the scans from native space to a subject specific space negatively affects the FreeSurfer segmentation, possibly by making the boundaries between tissue classes less clear. The longitudinal FreeSurfer pipeline recommends that segmentations should be performed on cross-sectional time points before registering them, however the large difference between the two analyses is concerning.

Both ANTs and MALP-EM showed more significant problems with measurement compared to FreeSurfer. ANTs segmentations were visually poor, with frequent overestimation of the occipital cortex and temporal lobes. This was more pronounced in atrophied brains, indicating a possible disease-related bias in the measurement of CT. In

addition, sulci that were tightly compressed were not delineated accurately, with the two sides appearing to merge, resulting in ultra-thick measures of the cortex at these points. The poor segmentations were reflected in the ICC and repeatability values. ICC values were generally between 0.70 and 0.80. While this is typically viewed as representing good reliability, the confidence intervals for ICC values for ANTs were very wide, and thus we cannot assume good reliability (Koo & Li 2016). In addition, repeatability values were >5 for most ANTs comparisons, again indicating poor reliability. Although scans were captured in one session, and thus we would expect to see little differences between measurements these values suggest that ANTs CT pipeline is difficult to apply reliably.

The results from the combination of MALP-EM segmented CGM and ANTs KellyKapowski CT pipeline, which calculates CT from previously segmented CGM regions, were more variable. Visual QC demonstrated that while overall the segmentations were of better quality than those produced by ANTs, issues were again seen with delineation in the sulci, similar to those seen with ANTs regions. Cross-sectional reliability indicated that MALP-EM plus ANTs CT pipeline was highly reliable. ICC values were >.95 for almost all regions, and repeatability was <2 for most regions. However, the longitudinal analysis indicated that both growth and reduction in CT of up to 42% for non-registered scans, and up to 25% for registered scans. Following analysis, post-hoc examination of some of the scans showing the most extreme reductions in CT were undertaken. Participants with large increases or decreases in thickness tended to have one CGM segmentation that was thicker by a number of voxels across the cortex, which could result in large differences between the two time (e.g. 25% increase in thickness in one registered scan pair). Some of the scan pairs showing large discrepancies often had one scan with reduced quality or difference in GM/WM contrast between the two scans. For scans showing large differences in thickness, the volumetric change calculated in chapter 3 were also re-examined and were not found to be disproportionately large (i.e. 1% reduction in volume over time for a scan showing a 25% change in CT). These large increases and decreases are likely to be artefacts of the technique and the sensitivity of CT to minor changes in scan quality, which have been noted previously (Clarkson et al. 2011), rather than real changes in thickness. Since voxel-wise techniques calculate thickness based on a value for every voxel in the cortex rather than every vertex in the cortical surface, they can be more heavily impacted by a consistent but minor over- or under-estimation of the cortex. Overall, the results support previous work indicating that volume-based methods are more susceptible to longitudinal instability than surface-based methods (Clarkson et al. 2011).

One of the most challenging aspects of calculating CT from MRI is creating highly accurate delineations of sulci, as described previously (Cardoso et al., 2011). Voxel-wise tools suffer from this problem due to the partial volume effect. If a very small sulcus consists of mainly GM from either side of the sulcus with a small amount of CSF, voxel-wise methods will classify the region as GM. Thus then quantifying CT from the voxel-map the sulcus will be missed, leading to very thick values for CT in this region. It is possible that this artefact explains the increased thickness values for both ANTs and MALP-EM compared to FreeSurfer. When comparing CT in healthy controls to a group with atrophy, such as HD, this artefact might artificially inflate group differences. As atrophy progresses and the sulci widen, the sulci might be classified more accurately due to increased CSF in the voxels, ultimately resulting in a reduction of overall thickness. This is demonstrated in Figure 4.8, which shows baseline and follow-up scans from a manifest HD participant in TRACK-HD, and demonstrates that as the sulci widen, more voxels are classed as CSF and thus CT measurement undergoes large reductions. While the widening of sulci is a disease-related effect, it is not a direct measure of CT and so can be considered an artefact rather than a true measure. This is a significant issue that surface based methods are less affected by as the delineation of sulci is also improved within FreeSurfer by the inclusion of a maximum and minimum value for thickness within the pipeline. That is, at no point in the cortex can thickness be >5mm when measured using FreeSurfer. Furthermore, the use of a minimum value may mean that the cortex is over-estimated in scans with extreme atrophy. While the maximum/minimum thickness values mean that that for some scans, thickness is likely to be either over- or under-estimated, it helps to ensure that the sulci are delineated clearly.

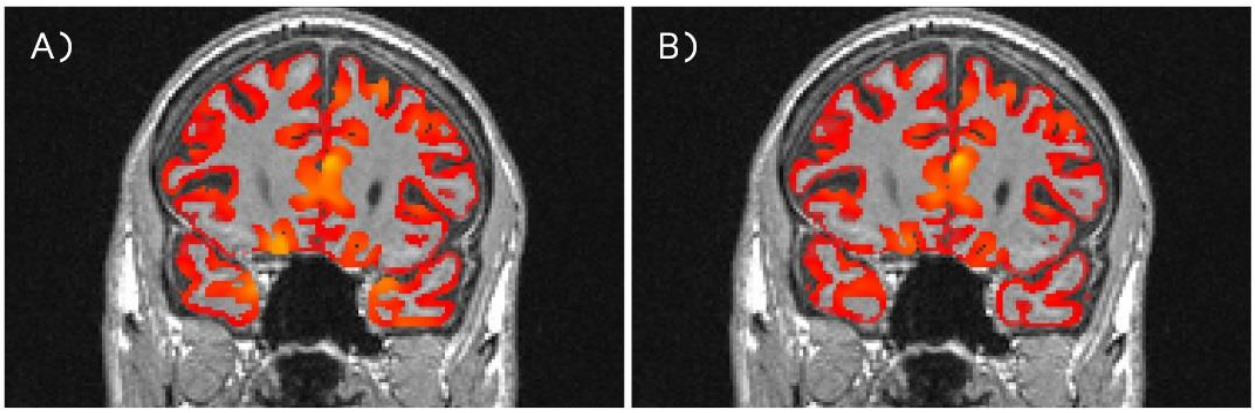


Figure 4.8. An example of a key source of variability in thickness measures. Figure (A) shows an HD participant at baseline, with Figure (B) showing the same participant three years later. In a number of regions, the sulci have widened with increasing atrophy, which causes a sharp reduction in thickness measurements at these points.

The large increases and decreases seen in some thickness measures over 3 years, along with the differing results found when running the analysis on both non-registered and registered scan pairs have implications for the measurement of longitudinal change in CT. It is important to note that in order to compare techniques more directly, the default FreeSurfer longitudinal pipeline was not used in this study. It is recommended that the pre-optimised longitudinal pipeline is used when running longitudinal analyses. However, internal evaluation comparing results from the cross-sectional FreeSurfer pipeline and the longitudinal pipeline has been performed previously on the TRACK-HD cohort, and the results were highly comparable and thus the decision was made to only use the cross-sectional pipeline in this analysis.

The results of this study support previous findings that have demonstrated the back-to-back reliability of FreeSurfer (Han et al. 2006; Tustison et al. 2014), and also one study which found that FreeSurfer produced more plausible longitudinal measures of change than voxel-wise methods (Clarkson et al. 2011). While these results provide further examples demonstrating the reliability of FreeSurfer, as measured quantitatively, this comparison also suggests that researchers should proceed with caution when using FreeSurfer to perform clinical comparisons, since the qualitative review of all scans demonstrated a high rate of poor quality thickness regions.

Only one recent study has performed an analysis validating similar voxel-wise measures to those included in this investigation. The previous study compared the reliability of volumetric measures to CT using similar tools in an Alzheimer's (AD) cohort (Schwarz et al. 2016). The study looked at a small number of regions related to AD pathology, measured the performance of FreeSurfer measures of volume and thickness, ANTs measures of thickness, and SPM volumetric measures and a combined SPM and ANTs thickness calculation, similar to the current application of MALP-EM and ANTs. Their results showed that thickness was generally less reliable than volume, significantly so in some regions, however thickness measures showed a stronger relationship to pathological measures of AD. Similar to the current analysis, which found high reliability for a combination of MALP-EM regions and ANTs thickness metrics, the authors found that the most reliable thickness measures were calculated using SPM segmentations and ANTs. They also found that the performance of ANTs thickness pipeline was variable, again, similarly to results reported here. While volumetric measures were more reliable, due to issues associated with the co-variation of head size that is a necessary requirement for volumetric analyses but not thickness analyses, the authors recommended that thickness should be used over volumetric measures. However, this study

only included cross-sectional analysis. Similarly, the results presented here indicated that cross-sectional regions had high reliability for the combined MALP-EM and ANTs pipeline. There was also a plausible trend for decreasing thickness with increasing disease burden. However, the previous study did not discuss the quality of segmentations other than to provide failure rates (classified as a gross failure), and they performed no longitudinal analysis. Longitudinally, the results of this analysis showed extreme variability in rates of change – indicating that this method is not suitable for longitudinal use, and also adding uncertainty to the validity of the cross-sectional findings.

4.5. Conclusions

This work has supported a number of previous findings, but also extended the literature by adding a longitudinal comparison for one relatively new measure of CT (ANTs) and one more experimental measure of CT (MALP-EM+ANTs). FreeSurfer and MALPEM+ANTs provided plausible cross-sectional findings, with MALP-EM+ANTs showing particularly high back-to-back reliability. Furthermore, FreeSurfer was sensitive to longitudinal change in non-registered scan pairs. However, a number of issues concerning the reliability and validity of these measures were also encountered. The accurate delineation of the cortical surface was a problem for all techniques, with voxel-wise measures suffering more from poor sulcal delineation, and the surface-based method, FreeSurfer, suffering from general segmentation inaccuracies. While these issues are not as important for cortical volume measurements, the accurate delineation of sulci makes a great difference to the calculation of thickness measures. Significant longitudinal change with increasing disease progression was only found when using FreeSurfer and only when scans were segmented in native space, with a completely different pattern of change found when looking at registered scans. When attempting to measure subtle cortical changes in a small number of participants it is essential that only the most accurate methods are applied. By including thickness measures, it is possible that spurious findings could be introduced due to artefacts of the techniques. Based on these results, the decision was made to exclude CT measures from the analysis conducted in Chapter 6. It is essential that the measures used to quantify CGM during motor onset are the most accurate and reliable measures, and this analysis has provided evidence suggesting that there are still issues with the use of CT measures.

5. VALIDATION OF AUTOMATED MEASUREMENT FOR SUBCORTICAL REGIONS

While cortical volume is the focus of the final analysis in chapter six, subcortical regions are well validated, highly sensitive markers of disease progression. The caudate and putamen show significant damage in post-mortem studies of HD, indicating the importance of these structures (Vonsattel et al. 1985). Furthermore, many studies have identified these regions as the earliest brain regions to undergo atrophy, as well as the regions that undergo the most rapid atrophy (Tabrizi et al. 2009; Tabrizi et al. 2012). The quantification of longitudinal change within both the caudate and putamen are two of the most robust biomarkers for observing change in HD gene carriers (Georgiou-Karistianis et al. 2013), as well as being strong predictors of disease progression (Tabrizi et al. 2013; Paulsen et al. 2010). Thus, it is important that they are included in investigations of disease progression for HD. They provide a useful measure of comparison, but also ensure that the group being studied are showing typical signs of HD neural changes. By including these measures, a unique understanding of the relationship between subcortical atrophy and cortical atrophy will be gained.

A number of HD studies have used validated manual measures of both the caudate and putamen to quantify cross-sectional and longitudinal volume in HD (Tabrizi et al. 2012; Tabrizi et al. 2009; Tabrizi et al. 2011; Tabrizi et al. 2013; Hobbs et al. 2013; Hobbs et al. 2015). In addition, the same measures are now being applied in multiple clinical trials for HD. While not an absolute GT, manual measures are often considered to be a 'gold standard' for segmentation since they are performed by highly trained raters with an understanding of neuroanatomical structure, following strictly defined procedures. However, these methods are time-consuming to perform and difficult for other research groups to replicate without training. Instead, automated measures of subcortical volume are often used in the literature to quantify subcortical volume (Paulsen et al. 2006; Domínguez et al. 2013; Aylward, Nopoulos, et al. 2011).

However, as with the segmentation of cortical regions, it is vital that the tools used to quantify volume are both reliable and accurate. The use of different tools to measure the volume of subcortical regions has resulted in some disparity between studies, with differences in atrophy rates for the caudate and putamen reported between different studies (Georgiou-Karistianis et al. 2013). Some studies have reported faster rates of atrophy in the caudate, and some in the putamen. These differences have implications for the selection of biomarkers in clinical trials, along with the sample size requirements for these trials (Georgiou-Karistianis et

al. 2013). While previous validations have been performed on subcortical regions from other tools (Perlaki et al. 2017; Rees 2014), MALP-EM subcortical regions have not been examined. After validation and subsequent selection of the MALP-EM cortical regions for inclusion in Chapter 6, the use of subcortical volumetric measures that have also been processed via MALP-EM would help to reduce inter-region bias introduced by the use of different measurement techniques to quantify volume in different regions. However, prior to doing this the subcortical regions produced by MALP-EM should be validated to ensure they are of a high standard.

5.1. Aims

This analysis aims to use previously segmented regions of the caudate and putamen to investigate the performance of MALP-EM on subcortical regions on an HD cohort. Scans from the PADDINGTON study underwent manual caudate and putamen segmentation at baseline, performed by experienced image analysts. The same scans were processed using MALP-EM, with qualitative examination of all regions and quantitative comparisons used to compare the volumes extracted from these regions.

5.2. Methods

Data for this study was taken from the PADDINGTON study (section 3.1.4). Data from the PADDINGTON study was used because all participants had both caudate and putamen regions manually segmented at baseline, providing a 'gold-standard' measure for caudate and putamen regions, whereas for TRACK-HD the putamen was not manually segmented.

5.2.1. Participants

PADDINGTON data from Leiden, London and Paris was included in this analysis. See section 2.1.4 for details on recruitment and data collection. These three sites also participated in the TRACK-HD and TrackOn-HD studies, with the same scanners and acquisition protocols. All participants who had a baseline scan which passed visual QC and had manual caudate and putamen regions were included in the study.

5.2.2. Segmentations

Manual segmentations of the caudate and putamen were performed upon data collection at the baseline time point of the PADDINGTON study, as described in section 2.3.3. Automated segmentation was performed in native space using the default parameters for MALP-EM. All regions underwent visual QC upon completion of the segmentation pipeline.

Manually measured TIV (section 2.3.3) was used to account for overall head size in group comparisons.

5.2.3. Statistical analysis

Summary statistics for both manual and automated methods were measured, and t-tests were performed to compare putamen and caudate volumes (as a % of TIV) between controls and HD participants. Intraclass correlations were also calculated to measure the consistency of manual and automated measures, with Bland-Altman plots and scatterplots used to visualise the relationship between manual and automated measures.

5.3. Results

5.3.1. Quality Control

Overall, the quality of the automated segmentations was high. The caudate regions were particularly good, especially given the large amount of atrophy in some scans. Some examples of manual and automated segmentations are shown in Figure 5.1 and Figure 5.2. However, there were a number of fails. One scan failed processing with MALP-EM due to incorrect classification of the GM and WM. The failed scan was examined and was found to be poorly position within the FOV. The scan was re-oriented and segmentation was attempted again. This time, processing completed successfully, indicating that the issue was associated with the initial registration step completed in MALP-EM. However, the scan was excluded from analysis as the manual segmentations were completed on the non-realigned data. In addition, there were two caudate segmentations and five putamen segmentations that failed due to underestimation of the region. An example of a failed caudate segmentation and a failed putamen segmentation are shown in Figure 5.3. These scans were excluded from the analysis.

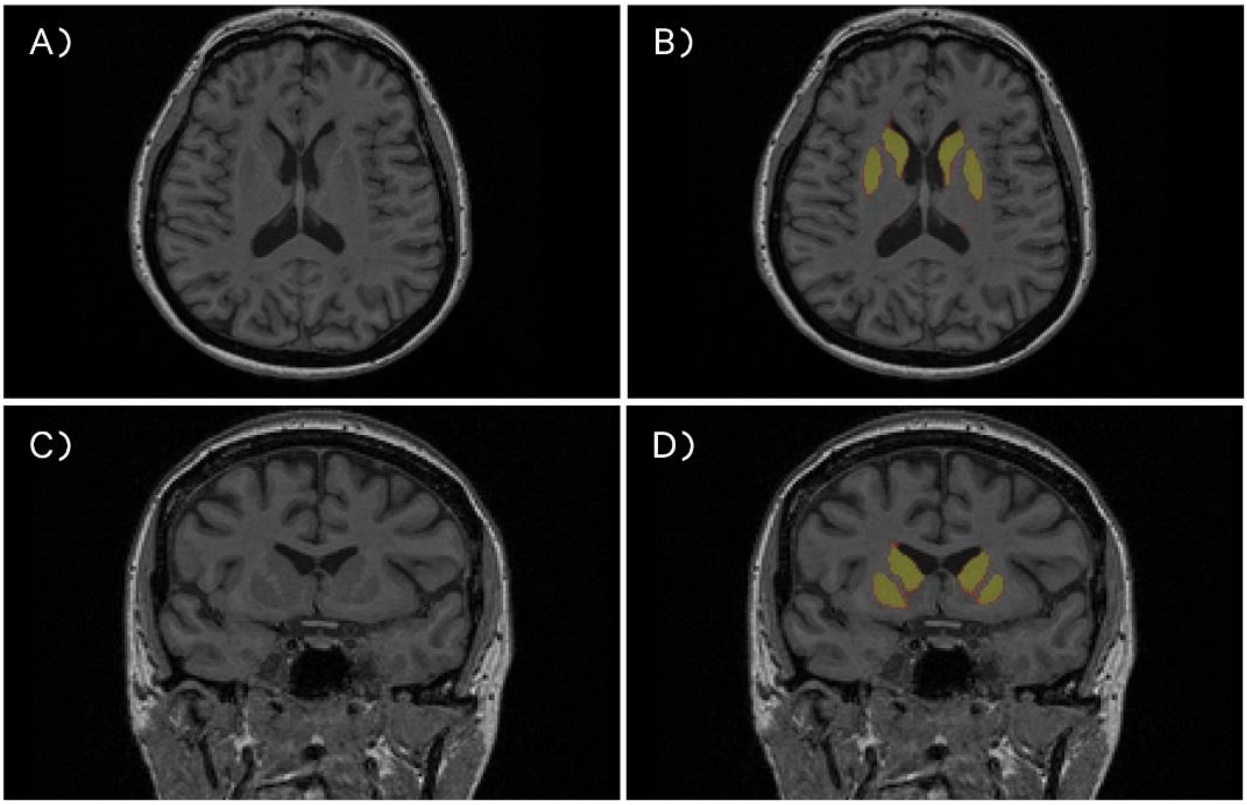


Figure 5.1. An example of automated measures of the caudate and putamen from the PADDINGTON study in (A) axial without a segmentation, (B) axial with a segmentation, (C) coronal without a segmentation, (D) with a segmentation.

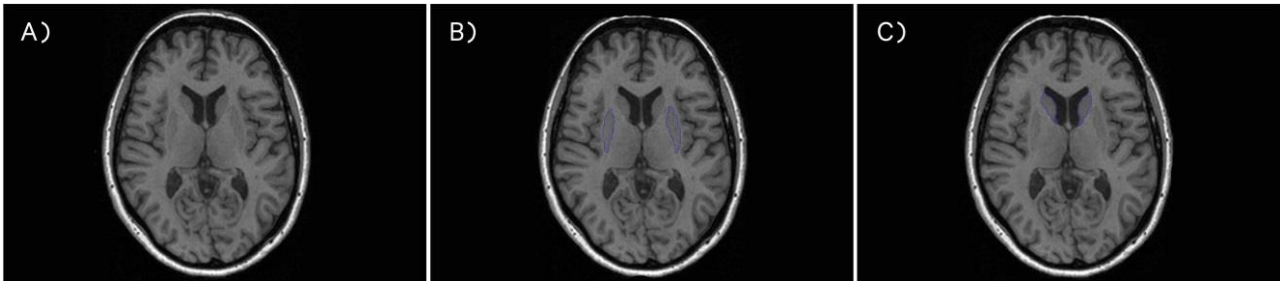


Figure 5.2. An example of manual caudate and putamen segmentations. (A) Shows a scan from the PADDINGTON cohort prior to segmentation, (B) shows the manually delineated putamen region and (C) shows the manually delineated caudate region.

5.3.2. Quantitative analysis

Participant demographics are shown in Table 5.1, and summary data for manual and automated regions is shown in Table 5.2. For both manual and automated techniques, HD participants had lower volumes of the caudate and putamen compared to controls (both raw volumes and as a percentage of TIV). Putamen and caudate volumes (% of TIV) were significantly lower for HD participants than controls for all measures at $p<.0001$.

Bland-Altman plots suggest no clear bias in measurement. Intraclass correlations were high, with consistency between putamen measures .943 (CI=.907-.965) and caudate measures .944 (CI=.909-.965). Finally, scatterplots (Figure 5.5) demonstrate a high level of association between manual and automated measures for both controls and HD participants. Scatterplots visualising the associations between manual and automated measures split by site were produced (Figure 5.6), and after visual inspection showing even distribution of data from the three sites, no further analysis was performed to examine the impact of site.

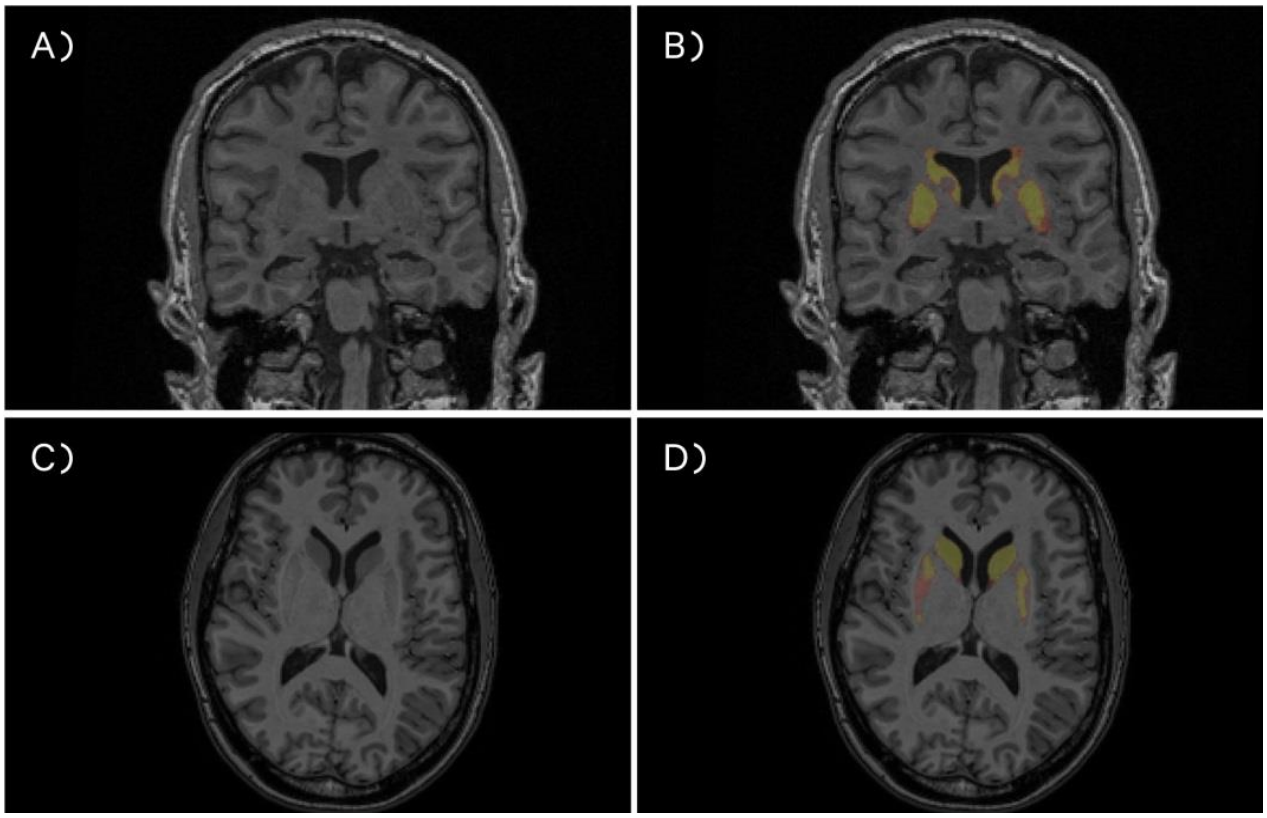


Figure 5.3. An example of failed automated segmentations. (A) Shows a coronal view without a segmentation, (B) shows the same view with a failed caudate segmentation, (C) shows an axial view without a segmentation, (D) shows the same view with a failed putamen segmentation.

Table 5.1. Demographics for the PADDINGTON participants included in the subcortical volume comparison.

	Controls (N=29)	HD (N=40)
Age	53.34 (7.60) 38.26-66.64	50.22 (9.99) 26.77-67.29
Sex (m/f)	12/17	9/31
CAG	-	43.35 (2.90) 39.00-54.00
Disease Burden Score	-	373.08 (87.75) 230.77-559.18
Total Functional Capacity	13.00 (0.00) 13.00-13.00	11.68 (1.27) 7.00-13.00
Total Motor Score	1.45 (2.03) 0.00-7.00	19.00 (10.38) 6.00-58.00

Table 5.2. Summary data and intraclass correlation for manual and automated putamen and caudate volumes expressed as raw volumes and as a % of TIV. t-tests showing the difference between groups for automated and manual measures (as a % of TIV) are also displayed.

		Controls	HD	Difference
Putamen	**Manual Raw**	7.75 (1.18) 5.17-10.70	4.97 (1.13) 3.03-7.11	-
	Manual % TIV	0.54 (0.07) 0.41-0.68	0.36 (0.08) 0.22-0.51	t(66)=9.83, p<.0001
	Automated Raw	7.59 (1.47) 4.83-11.44	4.75 (1.38) 2.47-7.68	-
	Automated % TIV	0.53 (0.09) 0.38-0.76	0.34 (0.10) 0.18-0.56	t(60)=7.77, p<.0001
Caudate	**Manual Raw**	7.49 (7.67) 6.12-9.11	4.99 (9.50) 3.33-7.30	-
	Manual % TIV	0.52 (0.05) 0.43-0.62	0.36 (0.07) 0.23-0.53	t(66)=10.63, p<.0001
	Automated Raw	6.44 (9.23) 5.15-8.22	4.28 (9.38) 2.84-6.42	-
	Automated % TIV	0.45 (0.05) 0.36-0.55	0.31 (0.06) 0.18-0.43	t(63)= 9.84, p<.0001

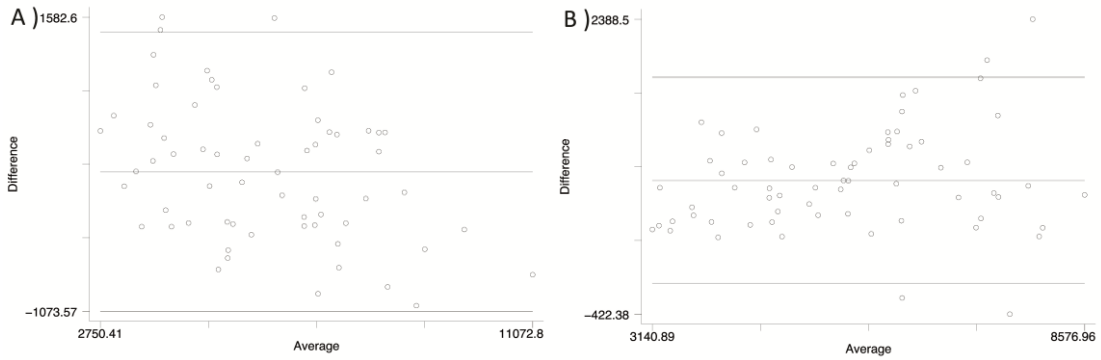


Figure 5.4. Bland-Altman plots comparing (A) manual and automated putamen volumes and (B) manual and automated caudate volumes.

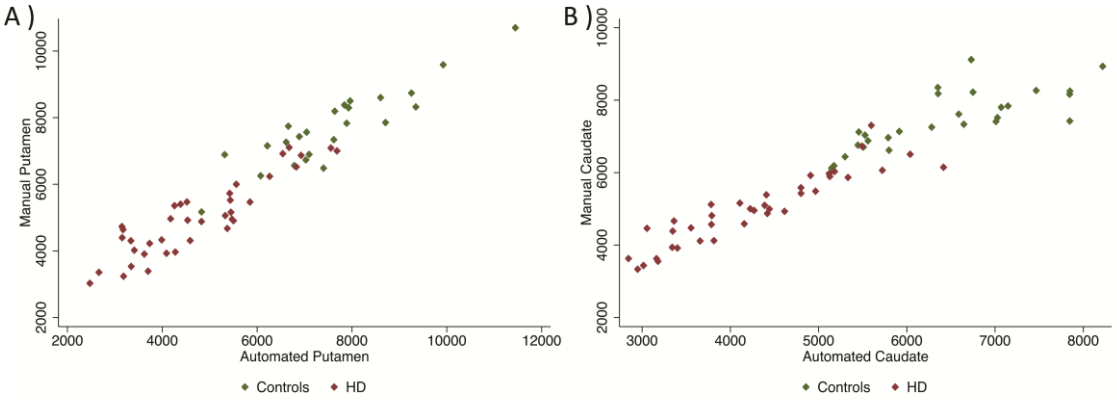


Figure 5.5. Scatterplots showing the relationship between manual and automated volumes of the (A) putamen and (B) caudate for both control and HD participants.

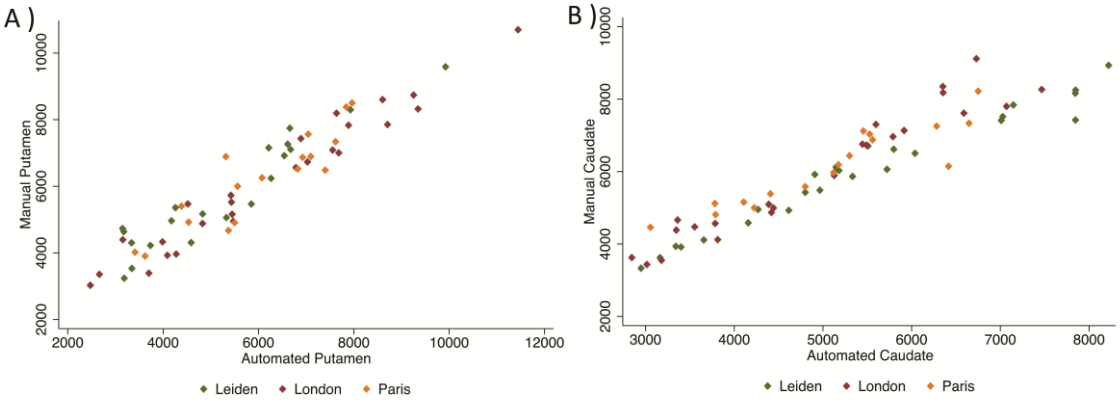


Figure 5.6. Scatterplots showing the relationship between manual and automated volumes of the (A) putamen and (B) caudate, for Leiden (green), London (red) and Paris (yellow) sites.

5.4. Discussion

This analysis aimed to validate an automated measure of the putamen and caudate quantified from T1 MRI scans. In Chapter 4, MALP-EM was validated as a sensitive and reliable

tool for measuring the volume of CGM in an HD cohort. Based on those results, it was selected as a candidate for measuring CGM in the longitudinal analysis to be completed in Chapter 6. However, it is important that subcortical GM structures are also included in the analysis for Chapter 6 as they provide a robust and sensitive measure of neural atrophy in HD. Ideally, the same technique would be used to measure both cortical and subcortical GM to reduce sources of bias when drawing conclusions about regional differences. However, subcortical volumetric measures extracted from MALP-EM have not been validated to-date. Here both putamen and caudate volumetric segmentations from MALP-EM have been successfully validated, demonstrating high correspondence with manual measures.

Similar to manual measures, the automatically segmented regions are highly effective at detecting significant group differences between control and HD participants in the PADDINGTON cohort. Automated measures of the caudate and putamen were significantly lower in HD than controls at $p<.0001$. In addition, very high levels of consistency were seen between manual and automated volumes, with intraclass correlations >0.94 for both regions.

However, it is important to recognise that a number of regions failed QC after being processed with MALP-EM. Two caudate regions had large holes that meant measurement would be incorrect, and six putamen regions were of poor quality, with underestimation of the putamen apparent. This failure rate could be problematic in small cohorts with limited data, and requires further examination in different cohorts. In the current data, the failures were not limited to either the control or HD group, suggesting that this is not an artefact stemming from atrophied regions. In addition, the poor segmentations did not appear to be as a result of a particular site or scanner type. This highlights the need for careful visual QC of segmentations from this pipeline prior to inclusion in future studies. While it is possible to perform manual editing on automatically generated regions to improve them, there should be strict criteria specifying when and how this should be done.

Based on the results of this analysis, automated volumetric measurements appear to perform to a standard approaching that of the 'gold-standard' manual measurements. While overall volumes were lower than manual regions, possibly due to the use of PVE on the automated regions but not on the manually delineated regions, there was a strong relationship between the two measures and automated measures showed highly significant relationships to disease stage. Automated measures of the putamen and caudate will be used in the following chapters of this thesis.

6. CHARACTERISING CORTICAL GREY MATTER ATROPHY DURING HUNTINGTON'S DISEASE MOTOR ONSET

Although many studies have detected significant atrophy within the cortex in pre- and manifest HD participants, previous research has been unable to present a thorough characterisation of the trajectory of cortical atrophy in HD because of limitations in study design and analysis methods that make assumptions about the linear progression of atrophy. It is more plausible to assume that neural changes are both linear and non-linear, with accelerations and decelerations possible during HD progression. Knowledge about the temporal progression of cortical change in HD could provide useful information to help understand the relationship between neural changes and HD symptomology. This chapter describes work that aims to expand upon previous findings by providing a detailed characterisation of CGM atrophy in HD participants using multivariate non-linear analysis to examine the progression of volume loss across multiple longitudinal time points. To provide a thorough characterisation of atrophy, linear modelling is used to detect regions that undergo atrophy at a constant rate (linear atrophy) and, additionally, non-linear models are used to track possible accelerations or decelerations in atrophy (non-linear atrophy). Previous studies have used cross-sectional data to make inferences about the progression of atrophy, but here, by modelling both linear and non-linear characteristics of atrophy in a carefully selected longitudinal cohort, the temporal progression of cortical atrophy can be better understood. This can in turn provide information about the impact of cortical atrophy on symptom onset and progression that could aid in the targeted development of therapeutic advances for the treatment of HD.

As described in section 1.1.6, at the point of end-stage disease cortical atrophy is widespread with only the temporal lobes relatively preserved. However, the nature of the progression of this widespread cortical atrophy during the pre- and early stages of manifest HD is still not well understood. A number of whole-brain analyses in HD gene-carriers suggest that volume loss begins in occipital and posterior regions and slowly progresses into more anterior regions of the brain as the disease progresses (Hobbs, Henley, et al. 2010; Tabrizi et al. 2012; Tabrizi et al. 2011; Tabrizi et al. 2009; Rosas et al. 2008). However, several studies have also indicated that frontal lobe volume is affected across both pre- and manifest disease stages (Aylward et al. 1998; Gómez-Ansón et al. 2009; Stoffers et al. 2010). A 2013 meta-analysis pooled the results of 17 VBM studies and found that between controls and pre-HD the only significant difference in cortical volume was in the occipital lobe, whereas for manifest HD,

differences were found in frontal cortex, primary motor, premotor and somatosensory regions, as well as in the intraparietal sulcus, the midcingulate cortex and the secondary somatosensory cortex (Dogan et al. 2013).

In an attempt to map the neural changes that occur in pre-HD, a small number of studies have used advanced analysis methods, as discussed in section 1.4.2 (Wu et al. 2017; Ciarochi et al. 2016; Coppen et al. 2016). To recap, two studies used cross-sectional data to infer the co-occurrence of regional change, thought to represent network-based atrophy. The results from Ciarochi et al. (2016) indicated co-occurring frontal and motor atrophy in HD gene-carriers far from onset, with parietal and occipital atrophy in those closer to onset. Coppen et al. (2016) found more widespread patterns of atrophy in pre-HD, but with similar occipital regions showing change in manifest HD participants. However, both studies used analysis techniques based on the assumption that atrophy both progresses linearly and is temporally-correlated between regions. It is more biologically plausible to assume that atrophy occurs in both linear and non-linear patterns, and that the rate and pace of atrophy differs across regions. Furthermore, methodological issues including the possible mis-registration of data in one of these studies, as described in 1.4.2, suggest that the findings should be interpreted with caution. Finally, a third study attempted to measure cortical change in a pre-HD cohort using cut-points - estimated points during disease progression whereby specific neural features begin to undergo significant change (Wu et al. 2017). Here, analysis including the WM, GM and subcortical structures did not show significant change within cortical GM, which may be due to the high variability in disease burden within the pre-HD cohort in addition to the use of stringent multiple comparison corrections applied due to the inclusion of multiple types of data. These studies were designed to characterise CGM change, and while they have expanded the findings of earlier VBM studies, there is still considerable inconsistency in terms of the pattern of cortical atrophy shown across the time course of HD and the true nature of CGM change in HD remains unknown.

In addition to the pattern of CGM atrophy during disease progression, there is also uncertainty about the effect of CAG on atrophy rate. Previous evidence suggests that CAG length contributes to both disease progression and atrophy rates (Rosenblatt et al. 2006; Henley et al. 2009; Hobbs, Henley, et al. 2010; Aylward, Mills, et al. 2011; Ruocco et al. 2008). However the exact nature of this effect is not understood, with a small number of studies indicating that higher CAG length results in faster atrophy of subcortical structures, as well as regions of the occipital, cingulate and frontal lobes (Henley et al. 2009; Hobbs, Henley, et al. 2010; Ruocco et al. 2008; Aylward, Mills, et al. 2011). The effects of CAG length on both linear

and non-linear atrophy have not been examined in detail. Further evidence describing this relationship may help to understand the progression of HD, particularly in high CAG individuals who tend to show faster clinical progression (Rosenblatt et al. 2006).

Collectively, the results of previous studies suggest that cortical atrophy generally occurs in late pre-HD and early HD, with the first atrophy likely to be occurring within occipital, motor and possibly frontal regions. However, the variability in terms of findings from these studies limits the conclusions that can be made about the progression of cortical atrophy in HD, particularly in pre-HD participants. There are a number of common methodological problems that may have contributed to this variability. Most previous studies use cross-sectional data to draw conclusions about the progression of atrophy, and often make the assumption that atrophy progresses linearly; as yet there are no studies that have quantified non-linear atrophy in a longitudinal cohort. The use of simple mass-univariate approaches means that limited conclusions can be drawn from these analyses (McIntosh & Misic 2013). The lack of work measuring non-linear atrophy is partly due to the complicated nature of the modelling required to perform analysis of non-linear atrophy across multiple time points. While complex modelling methods are widely available for functional and electrophysiological data, similar methods have not been commonly applied to sMRI data (Friston et al. 2003; Stephan et al. 2007; Friston et al. 2016). However, a recently published framework has been developed for analysing changes in brain structure over time using a dynamical systems method (Ziegler et al. 2017). Structural Dynamic Causal Modelling (sDCM) enables both linear and non-linear modelling of complex longitudinal data (Ziegler et al. 2017), with the option to apply variable input factors hypothesised to be driving neural change differentially within regions. Furthermore, change can be quantified within independent regions using a data-driven approach, but also within networks of regions to test specific network-based hypotheses about the causal links between atrophy in different regions. This framework was recently validated on pubescent brain changes, demonstrating that cortical growth during puberty progressed differentially between regions and that regional change was related to explicitly defined growth factors (Ziegler et al. 2017). Applying this approach to longitudinal data with multiple time points provides a powerful method of characterising structural change within the cortex in neurodegeneration.

A further limitation of previous HD studies investigating cortical atrophy is that they have applied a series of differing tools for the processing and analysis of MRI data. Recent studies have demonstrated that segmentation errors affect the conclusions drawn from volumetric neuroimaging studies (Johnson et al. 2017; Katuwal et al. 2016). When different

segmentation tools are used to process the same images, volumetric results can be significantly different and segmentations are often visually different between tools, with errors in segmentation thought to be a driving factor causing between-tool differences (Katuwal et al. 2016). These findings were supported by the results of Chapter 3, which demonstrated that inconsistency between tools can result in variable findings when applied on the same dataset. Due to the inconsistency introduced by the differing use of tools and substandard segmentations, it is essential that measures are validated as both reliable and accurate prior to use. Furthermore, visual QC should be performed on all scans and segmentations to ensure that any group differences or longitudinal changes are not just artefacts of poor segmentation techniques.

Another key issue associated with previous studies is the heterogeneity of HD symptoms and disease onset/progression across HD gene-carriers. Typically, studies examining neural atrophy in pre-HD use an estimated calculation of year to disease onset when grouping participants according to disease stage. This calculation is based on a combination of age and CAG repeat (Langbehn et al. 2004), and explains around 50-70% of variability in onset (Papoutsi et al. 2014). Although this formula can predict disease onset with moderate accuracy, when looking for subtle neural changes, heterogeneity within the groups can reduce the ability to detect small changes in the cortex and can result in variable findings across studies that have grouped participants differently (Paulsen, Long, Johnson, et al. 2014; Tabrizi et al. 2009). Given that a number of longitudinal HD studies have acquired imaging data within the same individuals spanning >6 years, it is now possible to select subsets of participants based on longitudinal characteristics such as speed of symptom progression (fast vs. slow progressors) or time of motor onset. This method of cohort selection can be used to reduce heterogeneity within groups, and thus increase the power to detect change within different phases of disease progression. However, retrospective grouping has only been used in a small number of studies to date and has not been used to study cortical change in detail (Paulsen, Long, Johnson, et al. 2014; Tabrizi et al. 2013).

Additionally, a series of widely-cited findings are based on data collected as part of the multi-site observational PREDICT-HD study, which includes MRI data from more than 30 sites, and both 1.5 and 3 Tesla scanners. Multi-site studies are important for advancing knowledge about HD and enable the recruitment of much larger cohorts; however, data collected from multiple sites may also introduce significant noise to later stages of analysis. While key measures that show large group differences or change in HD (e.g. caudate volume) are generally robust to the effects of using data from multiple sites, more subtle changes may not

be detected due to higher level of between-site variability. The inconsistency in previous findings for CGM atrophy in HD suggests that cortical change is slowly progressing and thus susceptible to the noise introduced by combining data collected from a large number of scanners with multiple field strengths, such that the subtle changes are hard to detect. Most previous studies have been cross-sectional, and using data from a large number of sites can be more problematic when performing cross-sectional analyses since between scanner differences in image quality can act as a significant confound for MRI analysis (Littmann et al. 2006; Focke et al. 2011). In contrast, longitudinal analyses involving the study of within-participant data are less affected by the variability introduced by multiple scanner types or acquisitions since participants are typically scanned at the same site with the same protocol throughout data collection and within-participant measures are used. Longitudinal data also accounts for natural variations seen in brain volume that can make group-differences difficult to detect in cross-sectional studies, and thus longitudinal analyses are preferable when possible.

6.1. Aims

In the following study, the aim was to characterise longitudinal linear and non-linear structural change within the cortex in HD, addressing a number of the shortfalls described above and providing the most detailed characterisation of the progression of CGM atrophy in HD to-date. This study uses tools validated in previous chapters to measure GM volume, and then applies a multivariate analysis method to quantify cortical GM atrophy in a sub-group of participants from the TRACK-HD and TrackOn-HD studies.

The cohort was selected according to motor symptom onset using a clinically rated measure of motor symptoms in HD. As described in 1.1.3, the diagnostic confidence score (DCS) quantifies motor symptoms, and is used to define the clinical onset of HD. This inclusion criterion was chosen for a number of reasons. Firstly, this period is a critical time in disease onset, with the increasing symptom severity an important marker of progression for both HD patients and clinicians. Despite the importance of this period in HD, there is little understanding about whether the increase in symptom severity is reflective of structural loss within the brain. By understanding the relationship between symptom onset and neural change, a greater understanding of the progression of HD pathology could be gained. In addition, given the heterogeneity of symptoms in HD, the use of DCS as an inclusion criterion for this study enables the investigation of neural atrophy during a stage of the disease that follows a more predictable pattern of progression and thus improves the likelihood of detecting homogenous group-wide cortical change.

sMRI data were segmented into cortical sub-regions using MALP-EM, which was validated in earlier chapters of this thesis; MALP-EM was also used to segment the caudate, putamen and global WM. The resulting cortical, striatal and WM segmentations were then modelled using sDCM to map the distribution and rate of cortical atrophy in HD participants in the decade surrounding clinical diagnosis. Four main analyses were performed. Firstly, to determine which regions underwent the highest overall volumetric reduction, linear models were applied to quantify total atrophy occurring over the 11 year period for each participant within each brain region. Then, the rate of linear atrophy within each region was measured over time to show which regions show the highest and most consistent rates of atrophy. The analysis was also performed to determine whether any regions of the cortex showed a pattern of non-linear atrophy, i.e., whether there was acceleration/deceleration of atrophy within cortical regions over the 11 year time period and if so, at which point during disease progression this acceleration occurred. In addition, since previous evidence suggests that CAG length contributes to disease progression and atrophy rates (Rosenblatt et al. 2006; Henley et al. 2009; Hobbs, Henley, et al. 2010) the effects of CAG length on both linear and non-linear atrophy were measured to determine whether CAG length contributes to atrophy rate. Finally, to investigate the association between behavioural change and cortical atrophy, a behavioural analysis was performed to map the progression of change in three performance measures, two of motor performance (one clinical, Total Motor Score, one behavioural, Speeded Tapping) and one measure of cognitive performance (SDMT). By studying the timing of change within these measures, the consequences of atrophy can be better understood.

6.2. Methods

6.2.1. Participants

Participants were from the TRACK-HD and TrackOn-HD study cohorts (see 2.1.2). For the current study, only participants who were initially included in the pre-HD group in both cohorts and subsequently transitioned to manifest HD ('converters') during the course of data collection were included. Participants were recruited during data collection for TRACK-HD or TrackOn-HD and categorised at recruitment as pre-HD based on the criteria described in 2.1.2. Pre-HD participants were included in the current analysis if they received a DCS of 4 at any subsequent time point, indicating that they had met clinical diagnostic criteria for manifest HD and had therefore converted to the manifest HD stage. Fifty participants met this criterion for conversion during TRACK-HD/TrackOn-HD. One additional participant received a DCS of 4 at

TRACK-HD visit 4, but then reverted to DCS <4 at a later time point and was excluded from this investigation.

All converters were re-aligned to consolidate year of motor conversion across all participants (Figure 6.1). This was done to increase homogeneity of disease progression within the group. Motor diagnosis is a disease progression milestone for HD, and thus by aligning the participants at time of motor-onset neural atrophy during the transition phase is predicted to be more consistent than looking across participants who are aligned differently. The first year of DCS = 4 was designated as year of conversion (time point =0), and each year prior to conversion labelled as year -1, -2, -3, etc. Every year after conversion was labelled as year 1, 2, 3, etc. This resulted in a spread of participants with data up to -6 years prior to motor diagnosis and 5 years after motor diagnosis, enabling the mapping of cortical changes occurring during HD conversion. See Figure 6.1 for a schematic showing the number of data points available for each time point after re-alignment of the data.

6.2.2. Measurement of motor and cognitive symptoms

In order to link neural changes to HD symptoms, three measures of disease progression were selected from the battery of TRACK-HD and TrackON-HD behavioural tests that have previously been shown to have strong associations with disease progression in HD (Tabrizi et al. 2013). Firstly, Total Motor Score (TMS) was used to approximate clinical motor progression (Huntington Study Group 1996). TMS measures the presence of motor symptoms, and ranges from 0-60 with a score of <5 indicating no substantial motor symptoms. Since TMS is a clinical measure and thus can be subject to rater bias, an additional measure of motor progression that is less subjective, speeded tapping variability in inter-onset interval in the non-dominant hand was also included. This measure requires participants to tap in time with a beep, and the tapping is measured. From this measurement, motor performance speed and accuracy can be assessed, and this measure shows significant disease-related change over 12 months (Tabrizi et al. 2011) and was predictive of decline on TFC after adjusting for age and CAG (Tabrizi et al. 2013). Additionally, the symbol digit modalities test (SDMT) was included as a measure of cognitive progression in HD (Smith 1991). The SDMT assesses visuomotor integration and has components of visual scanning. It is a pencil-and-paper task during which participants view a key showing symbols with the digits 1-9. The participants are then presented with symbols above empty box and are given 90 seconds to write the corresponding digits in the boxes. The SDMT has been demonstrated as a reliable measure for use in HD

studies, and is also related to HD progression (Tabrizi et al. 2013). It is a key cognitive measure used in HD studies.

6.2.3. Longitudinal image processing

Image processing involved a pipeline with three general steps: the structural data from all time points for each individual were registered together and an average scan was created for each participant; the average scans were then segmented into cortical and subcortical regions using MALP-EM; finally the regions for each participant were multiplied from the Jacobian map corresponding to each time point, creating a volumetric map for each participant at each time point and within each region. These steps are explained in more detail below.

For the first step, the longitudinal within-participant registration pipeline included within SPM12 (Ashburner & Ridgway 2012) was applied to each individual set of data, i.e. all time points for each individual. This process included between–scan registration, creation of an average scan (across the time points) and automated differential bias correction for between-scan inhomogeneity (Jenkinson et al. 2002). In using this pipeline, variability that is commonly seen within longitudinal measures of volume due to inconstant scan alignment or differences in inhomogeneity can be removed without introducing bias by selecting an image from one time point as a target for registration. The registration process was performed using default settings and all native space scans from each participant for each time point. The output included an average image (created by convolving the scans from all time points) as well as Jacobean images and warping parameters that were then used to align the native scans to the average scan. This allowed for QC on both the average image and the warped native image to ensure that registration had completed successfully for each time point. Registration initially failed for two participants due to poor initial alignment within the FOV between the scans from different time points. For these two participants, the native space scans were adjusted to be in rough alignment and registration was re-run successfully.

Following registration, MALP-EM was run on all average scans. MALP-EM is a freely available segmentation tool aimed at providing a fully automated method able to separate tissue classes in healthy brains as well as brains that show severe neural pathologies (Ledig et al. 2015). It utilises a previously described registration approach (Heckemann et al. 2012), atlas-based segmentation, and intensity-based expectation maximisation to segment the brain into cortical and subcortical regions (Ledig et al. 2015). As well as the validation performed as part of this thesis, MALP-EM has been validated on traumatic brain injury patients and with

successful stratification of patients into favourable vs unfavourable outcomes possible using MALP-EM segmentations (Ledig et al. 2015). MALP-EM outputs total volumes for different tissue types as well as regional volumes for 138 regions.

While the MALP-EM standard pipeline generally performs well on the TRACK-HD and TrackOn-HD data, minor regions of over-segmentation in the occipital and temporal lobe were occasionally seen (Chapter 3; Johnson et al. 2017). Thus to improve delineation of the cortex in these regions, manually segmented whole-brain regions created during initial processing for the TRACK-HD/TrackOn-HD studies were used as masks during segmentation to improve initial brain-extraction and reduce errors in these areas. The process of creating these regions is described previously (2.3.3.1; Freeborough, Fox, & Kitney, 1997), but, briefly, the procedure uses a semi-automated intensity-based threshold with manual edits to create a highly-accurate total-brain region in native space. These regions were created at baseline visits for TRACK-HD and TrackOn-HD participants, and were visually examined for accuracy prior to use in the current study. Minor edits were performed to improve these masks if necessary. The masks (in native space) were converted to NiFTI format, binarised and filled to ensure that the regions contained no holes. The mask was then registered to the average image for each participant using the warps generated during creation of the average image. Once overlaid on the average image for each participant, each mask was checked again to ensure that it overlaid the brain well. These masks were specified in the MALP-EM pipeline. All other settings for MALP-EM were the default values.

After processing, all segmented regions underwent visual QC to check for errors. One scan did not pass QC due to errors in segmentation that could not be rectified, but overall the standard of segmentations was high. Figure 6.2 shows a finished MALP-EM segmentation demonstrating the high standard of cortical delineation. For each participant, the regions segmented on the average scan were then multiplied by the Jacobean maps from each time point to create a volumetric map representing the volumes within each region at every time point.

MALP-EM automatically creates 96 default regions during segmentation, however a number of regions were combined to reduce the number of ROIs and corresponding noise within small cortical regions, and to ensure that results were biologically interpretable. There was a focus on an *a-priori* selection of 54 (27 bilateral pairs) grey matter regions of interest (ROI) and one global white matter volume. To create the reduced number of ROIs, default MALP-EM regions were combined. They were combined based on anatomical knowledge and visual inspection of the regions, with the regions that were combined and the final regions

shown in Table 6.1. As discussed previously, the regions were also combined across hemispheres. In order to facilitate more straightforward comparisons across regions the decision was made to analyse and present results of regional brain volume changes relative to volume at time point of motor diagnosis (in percent). Due to the temporal symmetry and the year of diagnosis being the only time point with full data availability in all subjects, this reference is expected to be more accurate compared to common referencing to volume at baseline scan.

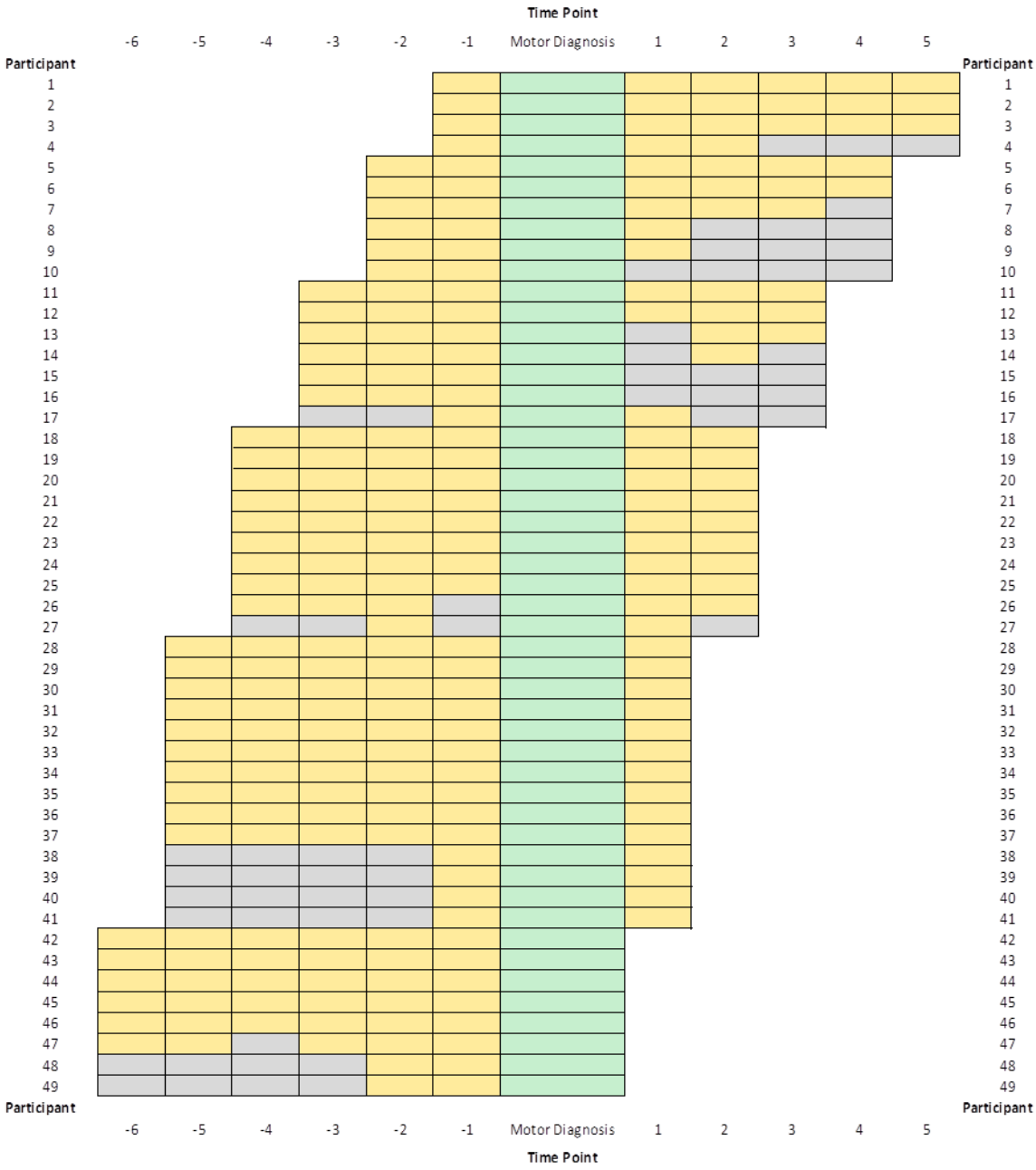


Figure 6.1. A schematic showing all data-points available for this analysis. Data is shown re-aligned so motor diagnosis is consistent between participants. Green represents motor diagnosis, yellow represents available MRI data and grey represents missing data. Missing data includes time points for which a participant was not yet recruited (e.g. when a participant was

recruited at baseline of TrackOn-HD, such as Participant 49), or had dropped out of the study (e.g. Participant 16 dropped out at the end of TRACK-HD and did not participate in TrackOn-HD), and when a participant could not attend a time point (e.g. Participant 26).

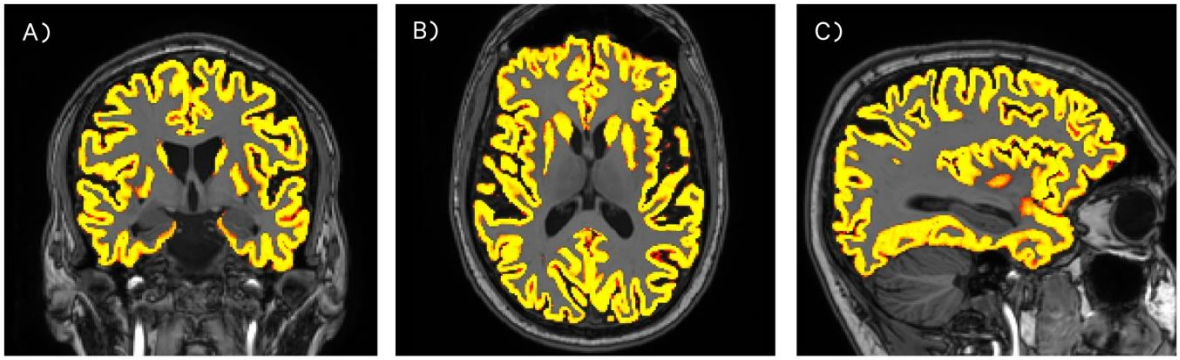


Figure 6.2. An example of a GM segmentation used in this study shown in (A) Coronal, (B) Axial and (C) Sagittal views. The figures show both cortical and subcortical regions.

Table 6.1. The final regions measured in this study, and the original regions output by MALP-EM that were combined to create the final regions.

Final Region Name	**Original Combined Regions**
Angular Gyrus Right	Angular Gyrus Right
Angular Gyrus Left	Angular Gyrus Left
Calcarine Cortex Right	Calcarine Cortex Right
Calcarine Cortex Left	Calcarine Cortex Left
Cuneus Right	Cuneus Right
Cuneus Left	Cuneus Left
Entorhinal Area Right	Entorhinal Area Right
Entorhinal Area Left	Entorhinal Area Left
Frontal Pole Right Occipital	Frontal Pole Right
Frontal Pole Left	Frontal Pole Left
Lingual Gyrus Right	Lingual Gyrus Right
Lingual Gyrus Left	Lingual Gyrus Left
Occipital Pole Right	Occipital Pole Right
Occipital Pole Left	Occipital Pole Left
Precuneus Right	Precuneus Right
Precuneus Left	Precuneus Left
Parahippocampal Gyrus Right	Parahippocampal Gyrus Right
Parahippocampal Gyrus Left	Parahippocampal Gyrus Left
Planumtemporale Right	Planumtemporale Right
Planumtemporale Left	Planumtemporale Left
Supplementary Motor Cortex Right	Supplementary Motor Cortex Right

Supplementary Motor Cortex Left	Supplementary Motor Cortex Left
Supramarginal Gyrus Right	Supramarginal Gyrus Right
Supramarginal Gyrus Left	Supramarginal Gyrus Left
Superiorparietal Lobule Right	Superiorparietal Lobule Right
Superiorparietal Lobule Left	Superiorparietal Lobule Left
Temporal Pole Right	Temporal Pole Right
Temporal Pole Left	Temporal Pole Left
Temporal Gyrus Right	Right Inferior Temporal Gyrus; Right Medial Temporal Gyrus; Right Planum Polar; Right Superior Temporal Gyrus; Right Transverse Temporal Gyrus
Temporal Gyrus Left	Left Inferior Temporal Gyrus; Left Medial Temporal Gyrus; Left Planum Polar; Left Superior Temporal Gyrus; Left Transverse Temporal Gyrus
Orbital Gyrus Right	Right Anteriororbital Gyrus; Right Gyrus Rectus; Right Lateral Orbital Gyrus; Right Medial Frontal Cortex; Right Medial Orbital Gyrus; Right Posterior Orbital Gyrus; Right Subcolossal Area
Orbital Gyrus Left	Left Anteriororbital Gyrus; Left Gyrus Rectus; Left Lateral Orbital Gyrus; Left Medial Frontal Cortex; Left Medial Orbital Gyrus; Left Posterior Orbital Gyrus; Left Subcolossal Area
Cingulate Gyrus Right	Right Anterior Cingulate Gyrus; Right Middle Cingulate Gyrus; Right Posterior Cingulate Gyrus
Cingulate Gyrus Left	Left Anterior Cingulate Gyrus; Left Middle Cingulate Gyrus; Left Posterior Cingulate Gyrus
Frontal Gyrus Right	Right Superior Frontal Gyrus; Right Superior Frontal Gyrus Medial Segment; Middle Frontal Gyrus
Frontal Gyrus Left	Left Superior Frontal Gyrus; Left Superior Frontal Gyrus Medial Segment; Middle Frontal Gyrus
Occipital Gyrus Right	Right Superior Occipital Gyrus; Right Inferior Occipital Gyrus; Right Middle OccipitalGyrus
Occipital Gyrus Left	Left Superior Occipital Gyrus; Left Inferior Occipital Gyrus; Left Middle OccipitalGyrus
Inferior Frontal Gyrus Right	Right Tringular Part Of The Inferior Frontal Gyrus; Right Orbital Part Of The Inferior Frontal Gyrus; Right Opercular Part Of The Inferior Frontal Gyrus
Inferior Frontal Gyrus Left	Left Tringular Part Of The Inferior Frontal Gyrus; Left Orbital Part Of The Inferior Frontal Gyrus; Left Opercular Part Of The Inferior Frontal Gyrus
Operculum Right	Right Central Operculum; Right Frontal Operculum; Right Parietal Operculum
Operculum Left	Left Central Operculum; Left Frontal Operculum; Left Parietal Operculum
Insula Right	Right Posterior Insular; Right Anterior Insula
Insula Left	Left Posterior Insular; Left Anterior Insula
Postcentral Gyrus Right	Post Central Gyrus Right; Right Postcentral Gyrus Medial Segment
Postcentral Gyrus Left	Post Central Gyrus Left; Left Postcentral Gyrus Medial Segment

Precentral Gyrus Right	Precentral Gyrus Right; Right Precentral Gyrus Medial Segment
Precentral Gyrus Left	Precentral Gyrus Left; Left Precentral Gyrus Medial Segment
Fusiform Gyrus Right	Right Fusiform; Right OccipitalFusiform Gyrus
Fusiform Gyrus Left	Left Fusiform; Left OccipitalFusiform Gyrus

6.2.4. Hierarchical disease progression models

The framework introduced in section 2.5 for longitudinal modelling of brain trajectories using dynamical systems is applied here (Ziegler et al. 2017). This modelling approach describes individual as well as group-level (hierarchical) change in regional brain volumes during the transition period from pre-HD to manifest HD. Details of the applied modelling and inference were previously published (Ziegler et al. 2017). Firstly, to map individual progression individual (first-) non-linear models are estimated. These individual models are then taken forward to the group (second-) level, where a model (the most appropriate for our data being selected using Bayesian Model Comparison; see below) is fitted that estimates change across the group whilst accounting for the effects of covariates (such as age, gender).

Individual first-level models were estimated using the established dynamical systems framework:

$$\frac{dx}{dt}(t) = Ax(t) + Cu(t, \theta_u)$$

The basic model assumes that $x(t)$ describes the change in regional volumes over time for each individual, ie27 bilateral volumes (25 cortical regions, as well as the caudate and putamen) and one global WM volume. The progression of x (volume in each region) is influenced by both endogenous dynamics, $Ax(t)$, and external inputs $u(t, \theta_u)$ (Figure 6.3. A). Internal dynamics (A) can denote regional self-connections and/or connections between regions. For this study, only regional self-connections are estimated representing within-region atrophy rates. Between-region interactions were not specified here due to a specific interest in examining change within all individual cortical regions prior to examining more specific, network-based hypotheses.

External inputs (u) are additional drivers of volumetric change or shrinkage, and in this model they represent the unknown underlying factors that may influence atrophy within a region. No explicitly defined external input factors, C, were included in the model.

This analysis has anticipated bilateral equivalence between the hemispheres and thus the volumes were combined across hemispheres, assuming that progression within each volume (x) describes the same evolution of volumes in both corresponding bilateral grey matter ROIs. This bilateral equivalence is based on previous sMRI results reporting largely symmetric effects of atrophy in the TRACK-HD cohort and in a meta-analysis of HD studies (Tabrizi et al. 2012; Minkova et al. 2017; Minkova et al. 2018).

Using this framework, the four analyses specified above were addressed to measure gross atrophy, linear rate of atrophy, non-linear accelerations in atrophy and the effects of CAG length on both linear and non-linear atrophy. The model specifications used to investigate these aims are detailed here.

6.2.4.1. First-level dynamical models of individuals

As part of the first or individual level analysis, a series of models were explored to identify which trajectory of volumetric change best explained the observed data. Seven different patterns of change including linear atrophy, as well as non-linear patterns that would imply acceleration or deceleration of atrophy during HD progression, each characterised by different external inputs, were modelled. The models were as following:

(a) Linear progression resulting in constant rates of atrophy trajectories;

(b) Piecewise linear progression assuming a global acceleration phase uniform across regions;

(c) Accelerations progressing in a sigmoidal (s shaped) manner. These accelerations can have regionally-specific sensitivity to change and rates of change, but with a global delay parameter common across all regions reflecting a delay in the acceleration of atrophy. The global delay parameter could be before or after motor diagnosis;

(d) Accelerations progressing in a sigmoidal (s shaped) manner. These accelerations can have regionally-specific sensitivity to change and rates of change, but this time with a regional delay parameter that reflects the possibility of a differing delay in the acceleration of atrophy across different regions. Again, the regional delay parameters could be before or after motor diagnosis;

(e) Volumetric changes follow a generalized logistic differential equation;

(f) Volume change follows a simple quadratic polynomial progression;

(g) Volumetric change evolves without the effects of external inputs.

DCM uses a probabilistic Bayesian framework and accordingly, assumptions are applied regarding the nature of estimated change in volume. Priors that reflect empirical or 'prior' knowledge of the distributions of expected volume change were therefore, included within the models, constraining the values of parameters during estimation. Here, only weakly informative priors representing data features were used rather than priors with strong assumptions about the progression of atrophy. This was because this is the first analysis of this type to be performed in HD, and it was unclear what pattern of degeneration might be found. Individual level model inversions were performed using previously established Variational Laplace methods (Friston et al. 2007). This was performed using the Variational Laplace algorithm for Bayesian parameter estimation, inference and model selection tools included n SPM12 via the SPM nlsi GN in release 6685 of SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm). More specifically, model inversion and estimation followed a symmetric system from time point of motor diagnosis (treated as the initial volume since data was available for all participants) both forward and backwards to the earliest time point prior to diagnosis and the latest time point after motor diagnosis. A detailed mathematical introduction to Variational Laplace and the applied inference using model evidence is provided in Ziegler et al. (2017). Model fitting can be potentially problematic, given that a good model may have a poor fit due to high levels of observation noise, while a complex model may have a good fit, but be biologically uninterpretable (known as overfitting; Stephan et al. 2010). Bayesian model selection guards against this problem by selecting a winning model that allows the optimal balance between accuracy and complexity (across the models being tested). It should be noted that while the winning model determines the precision of the estimates, it does not need to be considered explicitly in the later stages of analysis (Stephan et al. 2010). For this study, the Bayesian Model Selection is performed at the second level.

6.2.4.2. Second-level model of the group

Since the aim of this analysis is to perform group level disease progression modelling whilst accounting for individual level non-linear trajectories, each participant's first level model was embedded in a second (group-) level model. This enables the construction of group-wise models whilst controlling for individual level covariates, such as site, sex and age. A recently introduced empirical Bayes framework for estimation and inference of hierarchical non-linear models was used to estimate the group-level model (Friston et al. 2016). A number of participant-specific characteristics were included as covariates in the group-level to explain

first-level variability (Figure 6.3. C). These were: CAG length, sex, age (at motor diagnosis), TIV and scanning site. Notably, since CAG repeat length and age at diagnosis were highly correlated (r=-0.85), as is expected within HD studies, age was entered after orthogonalisation with respect to CAG.

This hierarchical modelling accounts for unexpected variation of first-level parameters, increases power for detecting group-level effects by accounting for differences in first-level parameters across participants, for example by accounting for participants with a differing number of visits, and explicitly allows for further examination of the effects of variables (e.g. CAG length) on all model parameters (e.g. atrophy rates).

Bayesian model selection was conducted comparing the obtained full hierarchical (two-level) models with all described first level forms and a second level including a group mean parameter and abovementioned individual covariates and confounds. Bayesian model evidence both optimises model fit while penalizing complexity and is therefore appropriate for model selection in highly parameterized hierarchical disease progression models (Penny 2012). Comparisons revealed that the evidence was highest for sigmoidal progression models with additional inputs to specify differential regional delays to the onset of accelerations (model (d) specified in section 6.2.4.1). The model evidence is shown in Figure 6.3 B. This model was subsequently used for the analysis. The regional inputs were specified for both the temporal delay of accelerations/decelerations and the sensitivity or amplitude of the accelerations. The winning model group results are presented in terms of the posterior distribution (expectation ± SD) of the group mean parameters.

6.2.4.3. Analysis of effects of CAG length on trajectories

As well as characterising overall atrophy, additional analyses were performed to examine the contribution of CAG length to the progression of atrophy in this cohort. The applied hierarchical disease progression model accounts for inter-subject heterogeneity by using covariates to predict first level parameters (Friston et al. 2016). Here, second level parameter effects of CAG repeat length were evaluated while accounting for all other covariates introduced in the first level. There was a particular focus on whether CAG effects were associated with differences in regional rates of atrophy or the amplitude and sensitivity to sigmoidal inputs causing accelerations or decelerations of atrophy. Furthermore, the effect of CAG length on absolute volume at point of motor diagnosis was measured using first level models with additional initial state parameters (volume at motor diagnosis).

Since this disease progression model is generative in nature, it allows individual trajectory predictions for all regions and participants over the whole decade around diagnosis. To further evaluate the relationship of genetic and brain differences as a process of the disease unfolding over years, all individual regional volumes were predicted using CAG repeat length for all time points independently while accounting for covariates and confounds.

6.2.4.4. Modelling volumetric change and behavioural results

The multivariate dynamical disease progression model was complemented by a separate model used to examine the change in motor and cognitive performance over time. This was done using a simple linear-mixed effects approach (based on Matlab function fitlme with Maximum Likelihood estimation). A quadratic model of disease progression (relative to time point of diagnosis) was fitted to the data for each behavioural variable independently. The quadratic fit was compared to more simple linear progression models using (simulated) likelihood ratio tests. The intercepts and slopes were allowed to vary across participants (as potential random effects) to account for the substantial heterogeneity in HD progression. The participant specific covariates of age at diagnosis, sex, CAG, and site were entered into the model as fixed effects. The models were used to determine whether performance on these variables accelerates by inspecting the linear and quadratic effects of time.

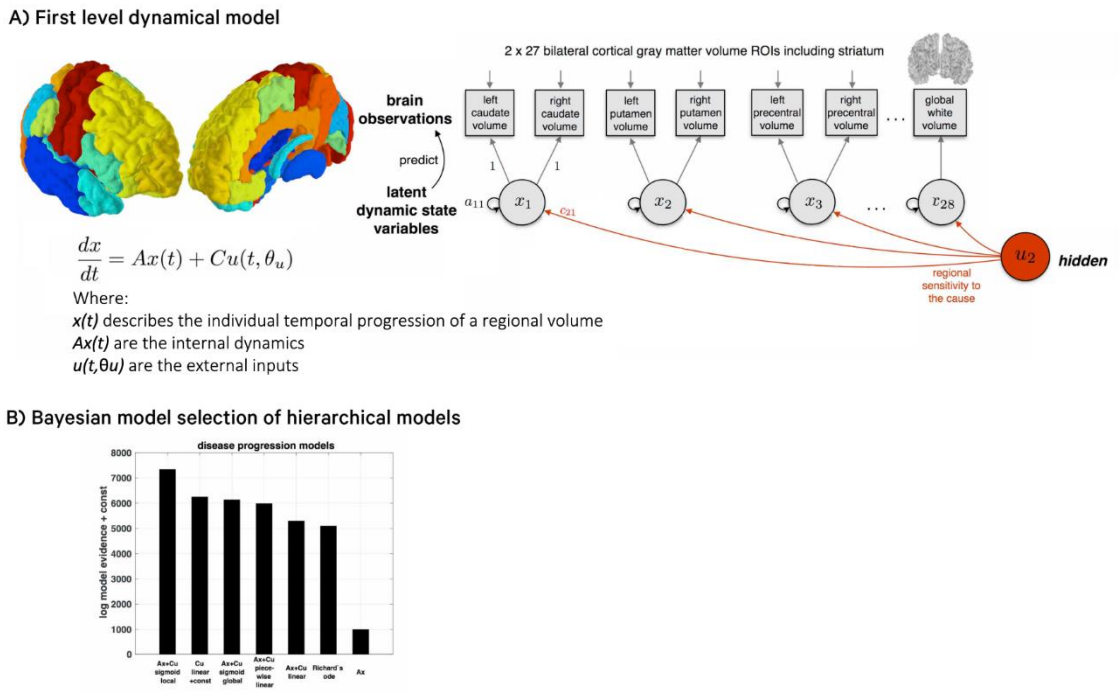


Figure 6.3. An example of the structural DCM model used in this chapter. (A) Shows the first level dynamical model, with the variables shown in the schematic on the right and (B) shows the log model evidence from the Bayesian Model Comparison for all models compared here, demonstrating that the sigmoidal model with regional delayed acceleration factors had the best fit.

6.3. Results

6.3.1. Demographics

Demographics for the final cohort are shown in Table 6.2. Demographics are the values measured at time point reflecting year of motor diagnosis.

6.3.2. Linear change during motor conversion

Linear modelling was used to quantify both total volume loss and rate of linear atrophy within all regions. Total volume loss (net atrophy) for the group is shown in Figure 6.4.A. Figure 6.4. A details the overall percent volume loss within each ROI, across the 11 years surrounding diagnosis for the group. During motor onset, there was higher volume loss in subcortical regions compared to that in cortical regions. Total atrophy for the putamen was 21.11%, and for the caudate was 19.68%, while within the cortex, motor and frontal regions displayed greatest overall atrophy, with the supplementary motor cortex undergoing the greatest change of 11.66% reduction in volume, followed by the frontal gyrus, precentral gyrus, superior parietal lobule, postcentral gyrus and inferior frontal gyrus, all showing between 8.52-10.46% change. Global WM underwent 8.25% change. Temporal lobes showed the lowest rates of atrophy, with the parahippocampal gyrus, temporal pole and entorhinal cortex showing particularly low levels of volumetric reduction over 10 years.

Following analysis of total volume loss, the rate of volume loss ('atrophy rate') across the 11 year period surrounding diagnosis was then examined, as shown in Figure 6.4.B. Atrophy rate represents approximate rate of linear change over the total time period surrounding diagnosis. Again, this rate was highest in the caudate and putamen, with frontal and motor regions showing the greatest rates of atrophy within the cortex. Temporal and occipital regions showed the lowest rates of linear atrophy.

Table 6.2. Demographics for the converters included in this study. The table shows mean (SD) and ranges, or N (%).

Age		44.59 (9.28) 28.65-66.00
Female		27 (55.10%)
CAG		43.67 (2.77) 39.00-50.00
Site	**Leiden**	22 (44.90%)
	London	10 (20.40%)
	Paris	10 (20.40%)
	Vancouver	7 (14.29%)
Total Motor Score		14.82 (5.63) 7.00-32.00
Speeded Tapping		-2.71 (0.52) -3.72- -1.49
Symbol Digit Modalities Task		44.00 (9.52) 22.00-65.00

6.3.3. Non-linear change during motor conversion

Next, sigmoidal progression models were used to model non-linear change, that is, the accelerations and decelerations of atrophy within each region. Additional inputs were included in the model to account for variation in the acceleration/deceleration (signal amplitude) of regional atrophy and the timing of these non-linear changes within the 11 year period surrounding disease onset.

Firstly, amplitude of non-linear volumetric change was accordingly calculated for each ROI, see Figure 6.5. Greatest acceleration of atrophy, i.e. highest amplitude, was seen in a number of motor regions, including the postcentral gyrus, the superior parietal lobule, the precentral gyrus and the supplementary motor cortex. Minor accelerations of atrophy were also seen in a number of other regions, including frontal and occipital regions and within the caudate and putamen. In contrast, minor decelerations of atrophy were shown in temporal regions and some medial-occipital regions, including the calcarine cortex.

For regions showing large accelerations/decelerations of change, timing of the accelerations/decelerations were estimated (Figure 6.5). It is important to note that the timing

of acceleration is limited to before, after, or around the time of diagnosis rather than a specific time point. As shown in Figure 6.5, the caudate, putamen and a number of occipital and frontal regions showed accelerated atrophy prior to disease onset. The largest acceleration of atrophy was seen in motor regions occurring post-diagnosis.

Figure 6.6. shows individual-level regional change and the corresponding sigmoid model for group-level acceleration of atrophy within each region. The plots demonstrate the relationship between individual-level change and group-level model fit, with variability seen between the regions. These plots demonstrate that some regions, including regions typically difficult to register and process such as the temporal lobes and parahippocampal region, show a high amount of noise. In contrast, a number of other regions, particularly subcortical, show less noise between time points. Figure 6.7 shows the individual model-change plotted alongside the group-model for one region, the caudate.

6.3.4. Effect of CAG on atrophy

Additional analysis using CAG as a second level parameter was performed to explore the link between CAG and cortical atrophy. Higher CAG length was associated with accelerated linear volume loss within occipital and subcortical regions, see Figure 6.8. In terms of non-linear change, those with higher CAG length displayed reduced acceleration of atrophy over the decade surrounding motor diagnosis, especially in sensory-motor regions; conversely, participants with lower CAG lengths showed greater acceleration of atrophy in these regions. Furthermore, participants with a higher CAG length had substantially lower caudate, putamen and WM volume than those with lower CAG length at time of onset, but higher volume within some frontal regions. Finally, higher CAG length explained increasing variability in volumetric change over time within subcortical and WM regions, but decreasing variability in a number of cortical regions.

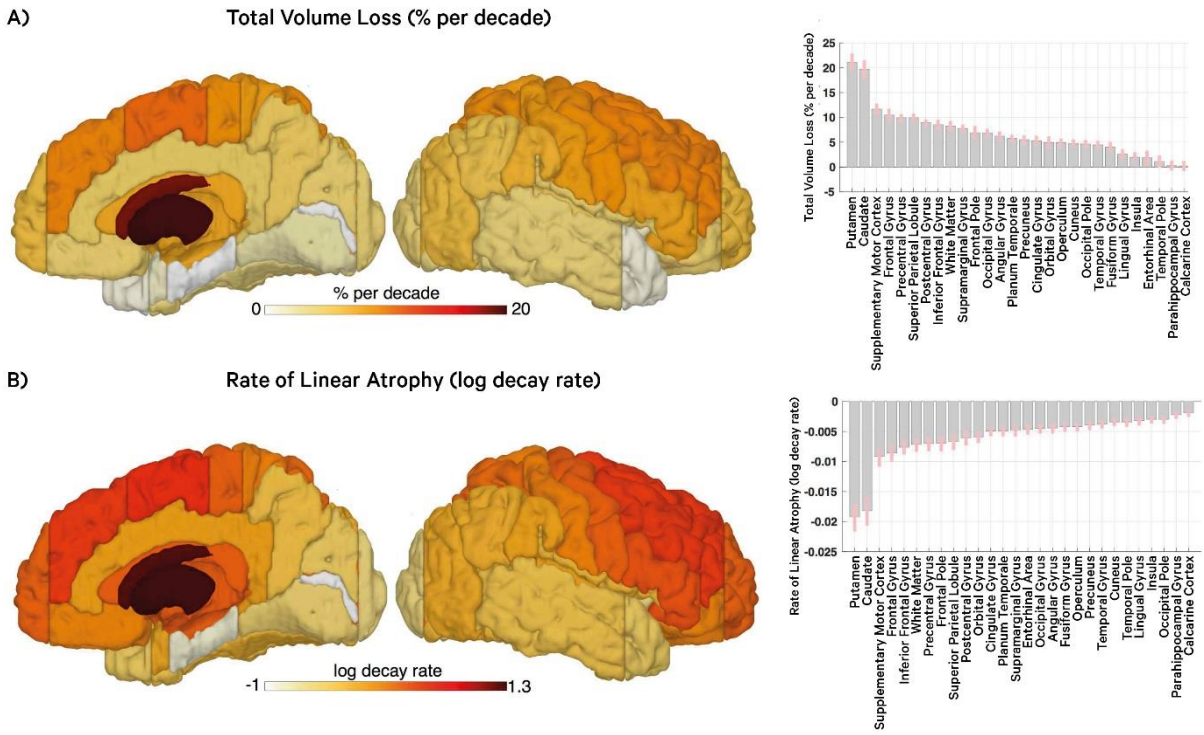


Figure 6.4. Group results showing (A) Surface projection (left panel) and parameter plot ± SD (right panel) of the overall percent volume loss of regional brain tissue per decade around diagnosis. (B) Rate of atrophy indicating approximately linear tissue atrophy during HD motor onset. These results account for the effects of age, sex, site, CAG and TIV.

6.3.5. Motor and cognitive change during motor conversion

Figures 6.9, 6.10 and 6.11 show the progression of TMS, Speeded Tapping and SDMT during the clinical transition from pre-HD to manifest HD in this cohort. For each figure, A shows individual-level model fit, and B shows group-level fit. Across the group, TMS began to increase thus representing an increase in motor symptoms before diagnosis which continued to accelerate after diagnosis. There were also minor increases in the Speeded Tapping score prior to diagnosis, indicating poorer motor performance; these increases in score accelerated after motor diagnosis. Finally, SDMT scores also began to show minor decreases across the group (indicating impairment in cognitive performance) prior to diagnosis, and similarly to Speeded Tapping, underwent accelerated decreases in performance across the group after diagnosis.

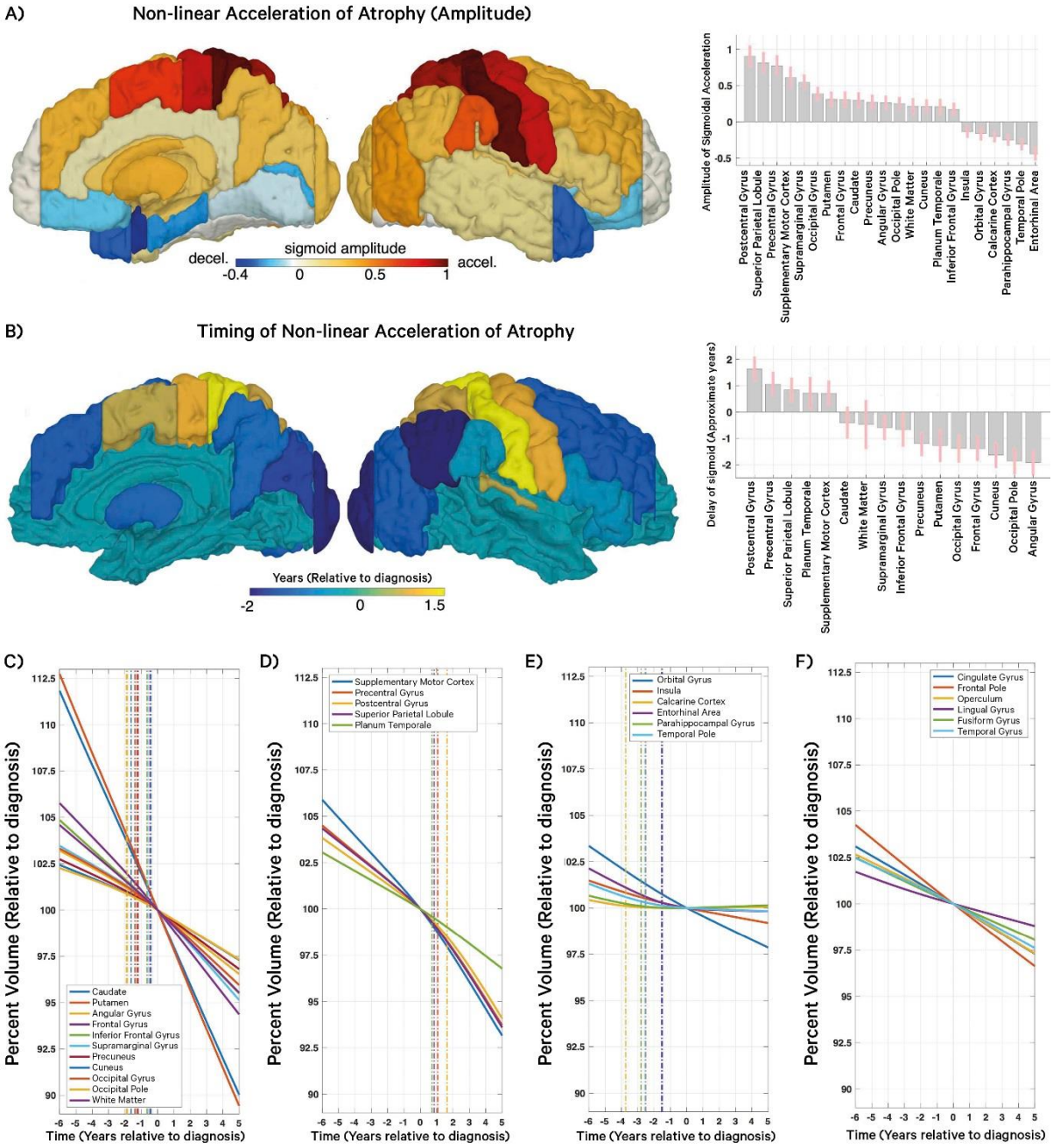


Figure 6.5. Group results showing (A) Regional sensitivity to input causing accelerations (red) or decelerations (blue) of atrophy (B) Delay parameters (of regions that show accelerated loss) showing approximate onset of acceleration. (C-F) Summary of all brain regions explored showing between-region differences in the group trajectory of atrophy progression. Regions differ qualitatively with evidence for some ROIs (C) showing accelerations before motor diagnosis (D) showing accelerations after motor diagnosis, (E) showing decelerations, (F) or no sign of non-linear effects (changes of rate of change). The disease model predicts individual and group level percent volume relative to volume at motor diagnosis accounting for individual variations in cag repeat length, age, sex, TIV and scanning site. Vertical dotted lines show the estimated time point of strongest accelerations of progression for each region.

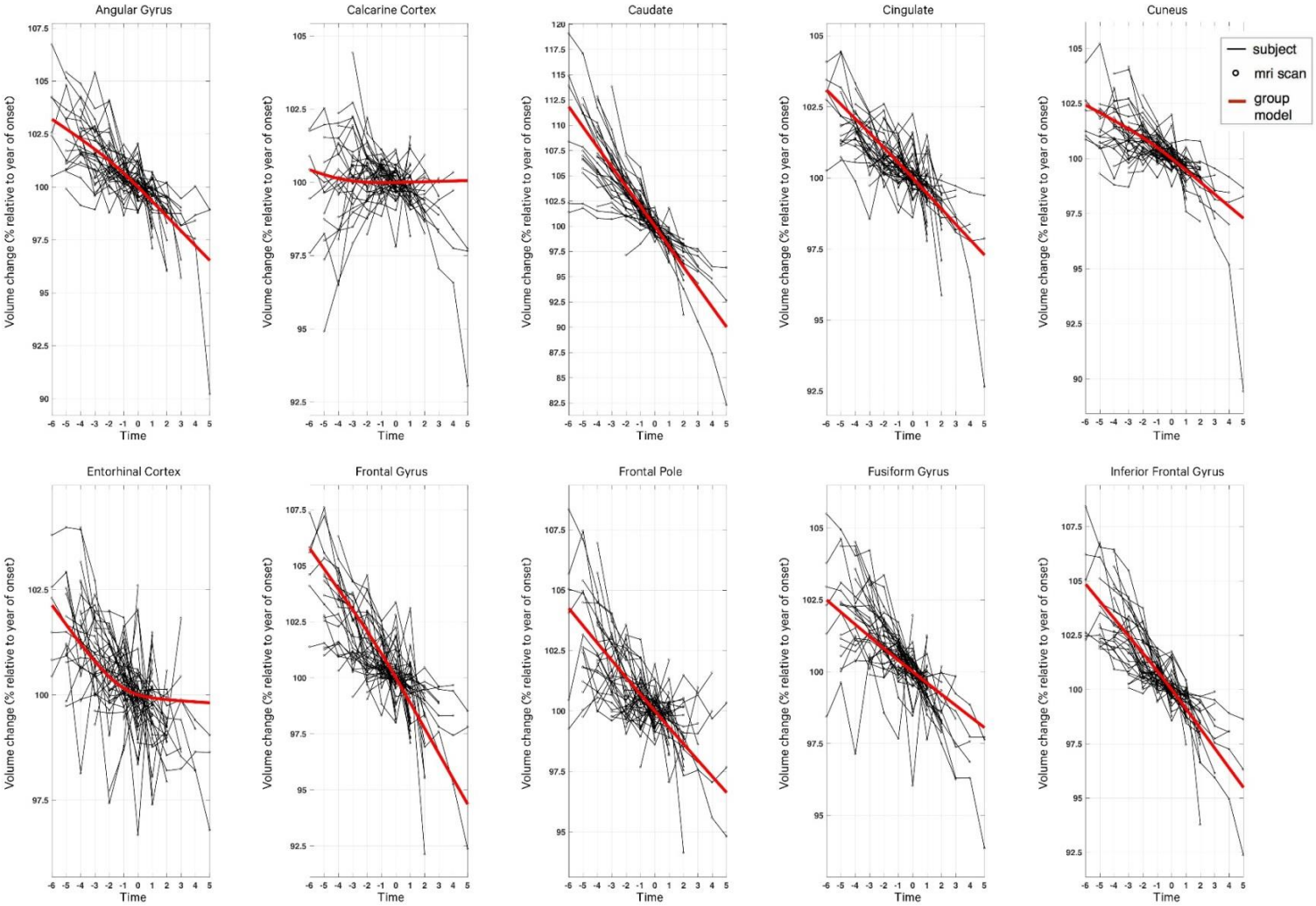
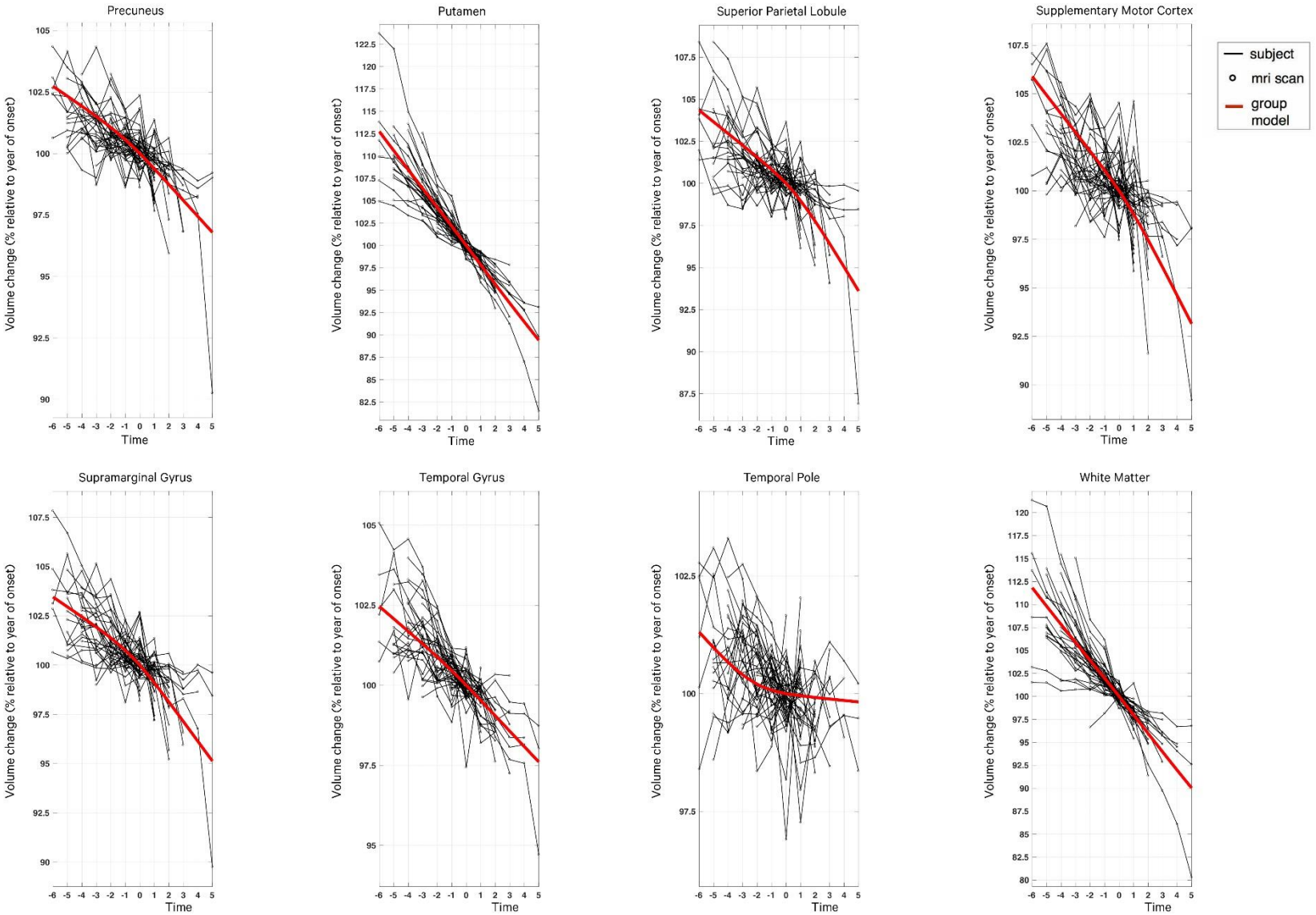


Figure 6.6. Examples of regional brain volumes and their progression during transition to HD. Plots show regional longitudinal raw data (percent volume relative to volume at year of motor diagnosis) with each black line representing one of the 49 participants. The group level dynamical disease progression model with highest Bayesian model evidence is shown in red. With exception of the white matter volume, regional volumes shown here refer to the left hemisphere with corresponding right hemispheric volume exhibiting very similar progression (not shown).

Insula
Lingual Gyrus
Occipital Gyrus
Occipital Pole
Operculum
Orbital Gyrus
Parahippocampul Gyrus
Planum Temporale
Postcentral Gyrus
Precentral Gyrus
Volume change (% relative to year of onset)
Time
subject
mri scan
group model


Precuneus
Putamen
Superior Parietal Lobule
Supplementary Motor Cortex
Supramarginal Gyrus
Temporal Gyrus
Temporal Pole
White Matter
Volume change (% relative to year of onset)
Time
subject
mri scan
group model

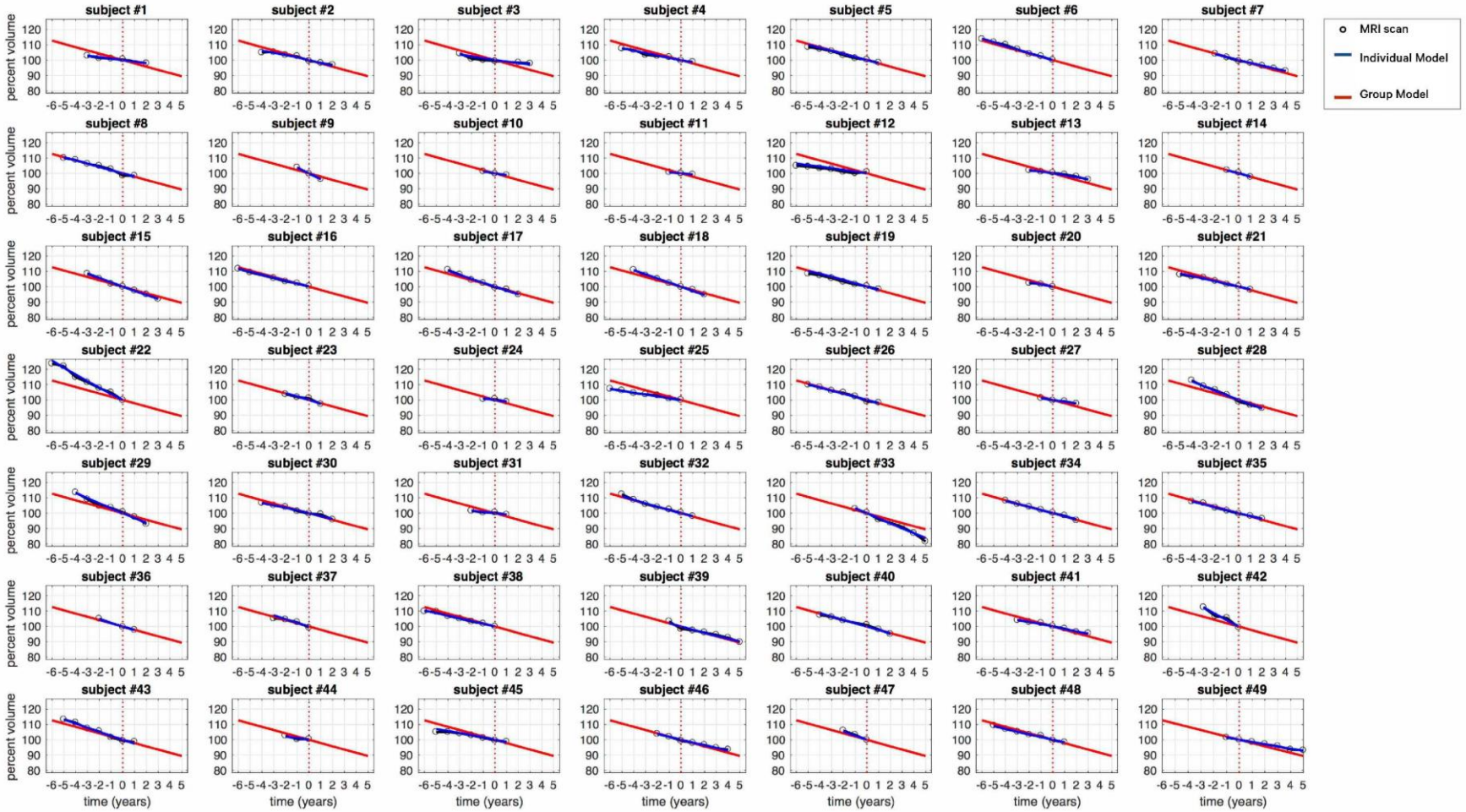


Figure 6.7. An example of the individual and group models, plotted for one region (caudate) to illustrate model fit.

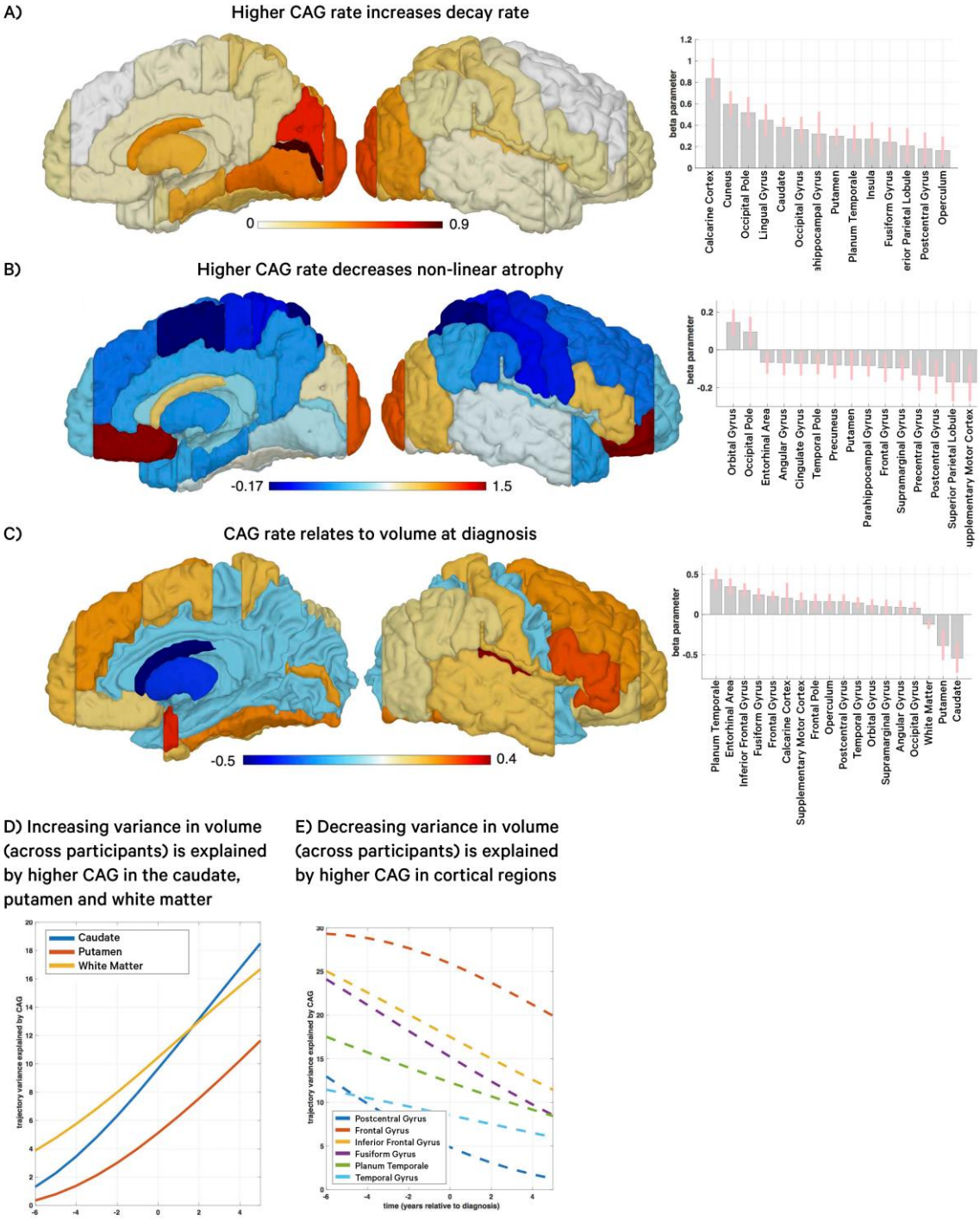


Figure 6.8. CAG repeat length is reflected in individual disease progression. (A) Higher CAG repeat length increases linear rate of atrophy in a number of regions. Shown is a brain surface projection (left panel) and parameter plot ± SD (right panel) indicating that CAG affects first level atrophy rate across patients (B) Higher CAG repeat length mainly associates to reduced amount of accelerations, rendering progression more linear in high CAG repeat individuals (C) Although generally younger, higher CAG participants have substantially reduced striatal and global white matter volume at time point of diagnosis. (D) Effect of CAG is also reflected in an increasing percent of variance explained in the putamen, caudate and white matter, and (E) a decreasing percent of variance explained in cortical areas.

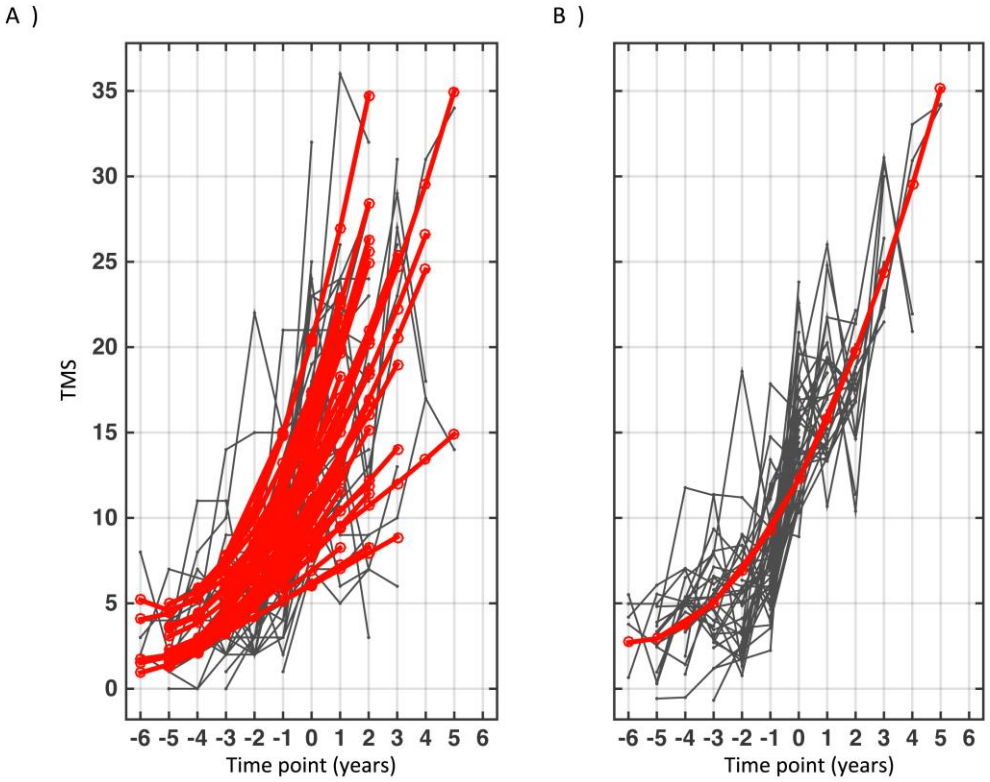


Figure 6.9. Graphs showing (A) individual-level and (B) group-level quadratic models representing change in Total Motor Score over time in all participants included in this study.

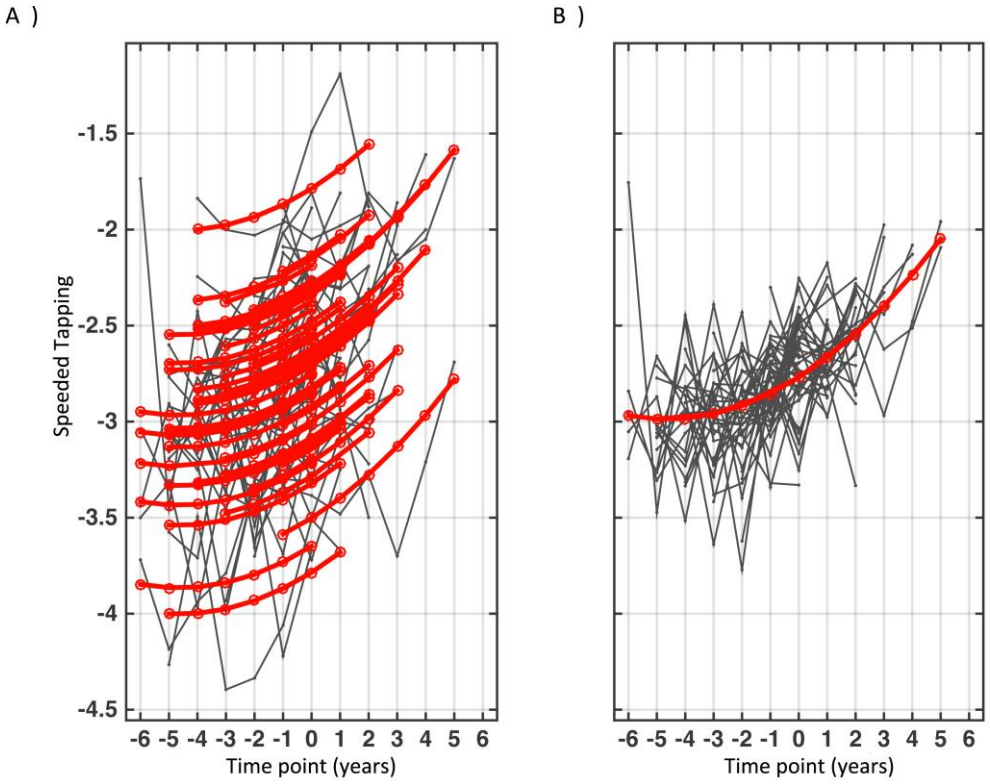


Figure 6.10 Graphs showing (A) individual-level and (B) group-level quadratic models representing change in Speeded Tapping over time in all participants included in this study.

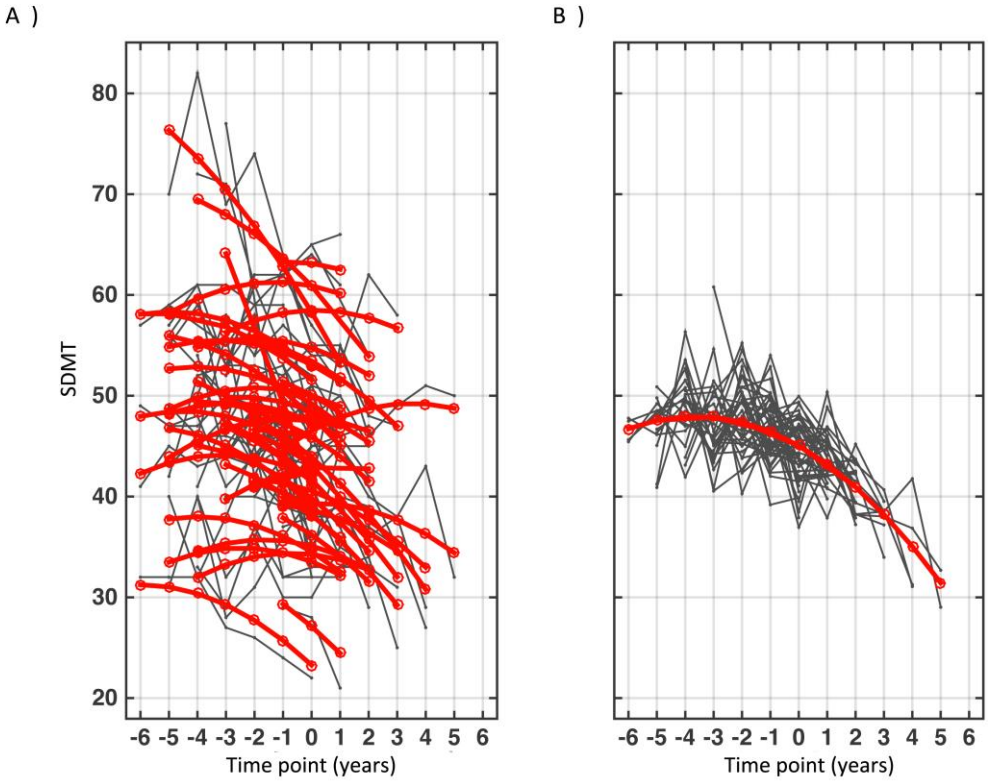


Figure 6.11. Graphs showing (A) individual-level and (B) group-level quadratic models representing change in SDMT over time in all participants included in this study.

6.4. Discussion

This study presents the first detailed mapping of linear and non-linear structural change in the cortex during conversion from pre-HD to manifest HD. Using a novel multivariate non-linear modelling approach it was shown that a number of cortical regions, particularly within the frontal and motor cortices, underwent significant linear atrophy during the decade encompassing clinical disease onset in HD. Concurrently, there was evidence of non-linear acceleration of atrophy mainly in motor regions that occurred in the period directly following clinical diagnosis. CAG length also had varying effects on linear and non-linear atrophy, with higher CAG lengths resulting in greater linear atrophy in subcortical and occipital regions, but smaller accelerations of non-linear atrophy in motor regions. A quadratic fit was used to map changes in motor and cognitive scores over time, indicating that TMS (a clinical motor score) shows accelerated impairment prior to motor diagnosis, whereas Speeded Tapping (a quantitative motor score) and SDMT (a cognitive score) begin to show accelerated deterioration after motor diagnosis. These findings are the first to highlight the differential patterns of atrophy that cortical regions undergo during motor onset, and have important implications for our understanding of the disease.

As expected, subcortical striatal regions showed greater overall volumetric reduction than cortical regions, with both the caudate and putamen showing total volumetric loss of around 20% in the decade surrounding diagnosis, but with the putamen showing slightly greater loss than the caudate. Subcortical striatal structures also showed the greatest linear atrophy rates and although there was some acceleration of atrophy that occurred around the time of diagnosis, overall subcortical atrophy showed a predominantly linear pattern. This result supports previous work suggesting that subcortical regions appear to undergo atrophy that follows a linear pattern beginning long before motor onset (Hobbs, Barnes, et al. 2010). Furthermore, the results provide evidence in support of marginally faster atrophy of the putamen than the caudate, similarly to previous TRACK-HD studies (Tabrizi et al. 2012). There have been contrasting findings in the literature when comparing rates of caudate and putamen atrophy (Georgiou-Karistianis et al. 2013), with differences between studies attributed to the varying use of segmentation methods and different cohorts (pre-HD vs manifest HD and different datasets). While this study does not provide a definitive answer as to whether the caudate or putamen undergoes faster atrophy, and it is possible that rates differ over the course of the disease, the results indicate that during transition to manifest HD the putamen is undergoing a slightly faster rate of atrophy.

For the first time, this study has demonstrated that a number of frontal and motor regions undergo the greatest total cortical atrophy and the highest linear rates of atrophy during conversion to manifest HD (Figure 6.4.). Several regions lost between 8.25-11.66% of baseline volume over the 11 year period. While previous work has demonstrated that motor and frontal regions are significantly reduced in pre- and manifest-HD compared to control participants (Thieben et al. 2002; Gómez-Ansón et al. 2009; Sormani et al. 2004; Wolf et al. 2009; Gavazzi et al. 2007; Aylward et al. 1998; Tabrizi et al. 2009), the present results provide detailed regional rates of progression. The supplementary motor cortex showed the greatest net atrophy in this cohort, with the frontal gyrus, precentral gyrus, superior parietal lobule, postcentral gyrus and inferior frontal gyrus indicating a distinct pattern of motor and frontal change. Furthermore, similar frontal and motor regions also showed a high rate of linear atrophy; with frontal regions generally displaying slightly higher rates of linear atrophy than motor. Interestingly, both the net atrophy and linear rates of change within the temporal lobes indicate that they are relatively spared from degeneration during this time period; a finding that is supported by both in-vivo and post-mortem results (Douaud et al. 2006; Tabrizi et al. 2009).

Within some regions, the total amount of atrophy occurring over the 11 years measured in this study was much higher than annualized rates typically reported from measures of global GM volume in both pre-HD and manifest participants, which has been quantified at around 0.5% per year in manifest HD2 participants, and lower in HD1 and pre-HD participants (Tabrizi et al. 2012). This is unsurprising when considering the variability of atrophy seen within the cortex. Regions undergoing little change, such as the temporal lobe, may be diluting the effect of regions undergoing significant atrophy in global GM measures, even within manifest HD participants. Thus, previous work quantifying global GM change may have under-estimated cortical change and contributed to a view that cortical atrophy in HD has little effect on disease progression in comparison to subcortical and WM changes.

This study also provides novel insight into more complex non-linear patterns of atrophy across time, seen in Figure 6.5. Motor regions underwent a striking pattern of acceleration of atrophy during transition from pre-HD to manifest HD, with further analysis demonstrating that this acceleration occurred shortly after clinical motor diagnosis. Non-linear analysis of CGM has not previously been performed in HD, and while this finding is perhaps unsurprising as clinical diagnosis occurs at the point of motor conversion, it provides novel evidence demonstrating the significant acceleration of atrophy within motor regions over the period of conversion. It is interesting that atrophy undergoes acceleration shortly after diagnosis, since motor diagnosis occurs at the point in disease progression whereby motor symptoms are sufficiently evident to lead to diagnosis and therefore higher acceleration of motor cortical atrophy may be expected prior to the point of conversion. However, it seems likely that degeneration in the white matter precedes that of cortical grey matter (Tabrizi et al. 2009; Paulsen et al. 2008; Wu et al. 2017). As such, the white matter pathways that connect cortical grey matter regions could be primarily affected prior to diagnosis; with the atrophy in the cortical grey regions occurring later in the disease trajectory of neurodegeneration. Thus, it is the changes in white matter or structural connectivity rather than cortical atrophy that drive symptom progression. McColgan (2017) demonstrated significant reduction in connection strength in WM connections to posterior regions, including motor regions, over 24 months in pre-HD participants. Furthermore, the significant change-points seen in WM volumetric and diffusion metrics but not CGM volume across pre-HD participants in the analysis comparing volumetric and diffusion metrics by Wu et al. (Wu et al. 2017) provide further support for this theory. Future work should aim to directly address this hypothesis. In this study, the WM underwent around an 8% decrease in volume over the period, and showed a predominantly linear pattern of atrophy with only minor acceleration in atrophy seen shortly after diagnosis.

Previous studies have reported lower cross-sectional WM volume in both pre-HD (Aylward, Nopoulos, et al. 2011; Paulsen et al. 2010; Tabrizi et al. 2009) and manifest HD compared to controls (Aylward et al. 1998; Halliday et al. 1998; Tabrizi et al. 2009). Longitudinal studies have also shown significant change in WM in pre-HD and manifest HD (Aylward, Nopoulos, et al. 2011; Ruocco et al. 2008; Tabrizi et al. 2011; Hobbs, Henley, et al. 2010), with rates of around 1% WM loss per year in pre-HD and around 2% per year in manifest HD (Tabrizi et al. 2012), which are in line with the current rates of change. However, it is important to recognize that global WM atrophy may not be a true marker of WM progression in HD and that DTI or connectivity metrics better reflect WM properties. Because of this, detailed conclusions on the rate of WM atrophy cannot be made based on the current results.

Some frontal, lateral occipital and both striatal regions showed minor acceleration of atrophy occurring prior to disease onset, yet these accelerations were much less pronounced than those seen in motor regions. The minor accelerations in frontal regions that precede those seen in motor regions could help to understand the onset of cognitive disturbance that often occurs prior to motor onset in HD, and this could be investigated in a cohort further from onset. Finally, temporal and some medial-occipital regions underwent slight deceleration of atrophy, however these regions showed lower rates of atrophy and deceleration was less pronounced. In addition, as shown in Figure 6.6., some of these regions were particularly noisy regions.

These results reveal an interesting pattern of both linear and non-linear cortical change. For the first time, this analysis has highlighted the degree to which sub-regions of the cortex are affected differentially by HD motor onset. Both similarities and differences in progression can be seen between functionally distinct regions of the cortex, for example frontal and motor regions undergo similar rates of linear change, but motor regions show a much greater acceleration of atrophy on top of the consistent linear progression seen. Regions including the supplementary motor, pre- and post-central gyrus underwent high rates of linear atrophy as well as strong accelerations of non-linear atrophy, as shown in Figure 6.4 and Figure 6.5. In contrast, while the frontal lobe also appeared to undergo significant linear decline, it was not as affected by accelerations in atrophy, with more minor accelerations seen prior to diagnosis. These results highlight the complex trajectory of CGM change in HD, and provide further evidence that previous methods of analysis are unsuitable for characterising this change. They also support the work of McColgan et al. (2015; 2017), who showed that WM connections between the striatum and cortical regions in the parietal and frontal regions appear to deteriorate earlier than other WM connections across the brain in HD. McColgan

theorised that early degeneration may be occurring in cortico-striatal tracts that are the most highly connected and experiencing the most neural traffic. These are also tracts connecting the basal ganglia to regions within the motor and association networks (Section 1.5). In the current study, the cortical regions that showed the greatest atrophy were the frontal and parietal regions, supporting McColgan's work and indicating that atrophy is occurring within regions of the motor and association networks during HD motor conversion. It is currently unclear why these two networks are particularly susceptible to degeneration, and why degeneration appears to spread from the basal ganglia towards cortical networks via the WM tracts. Future work should be done to further link degeneration within the WM and GM of these networks.

The effect of CAG length on atrophy was also modelled here. Previous work has demonstrated a link between CAG length and HD progression (Rosenblatt et al. 2012; Tabrizi et al. 2013) with higher CAG associated with faster HD progression. The results indicate that higher CAG lengths are associated with higher rates of linear atrophy within the caudate, putamen and across the occipital lobe (Figure 6.8.). This supports the findings of a number of previous studies relating higher CAG repeats to greater linear change within the subcortical and occipital lobes (Hobbs, Henley, et al. 2010; Ruocco et al. 2008; Henley et al. 2009). Finally, higher CAG participants showed lower volume of the caudate, putamen and WM at diagnosis than those with lower CAG repeats, although they also had higher volume within a number of frontal regions. It could be interpreted from this result that caudate, putamen and WM are more severely affected than cortical regions by a higher number of CAG repeats. It is also possible that because of the strong association between age and CAG length meaning that converters with high CAG repeats were generally younger, frontal regions may be showing a protective effect of age. Interestingly, occipital regions typically found to undergo substantial atrophy in both pre-HD and manifest-HD showed lower atrophy and rates of atrophy in this analysis, despite being named in numerous studies as one of the earliest regions to undergo atrophy in HD (Coppen et al. 2016; Hobbs, Henley, et al. 2010; Tabrizi et al. 2009; Tabrizi et al. 2011). While the occipital lobe did undergo significant atrophy in this study, it was not as prominent as previous results suggest. However, the strong association seen between CAG length and atrophy in this supplemental analysis suggests that rate of occipital atrophy is closely linked to CAG length (which was controlled for in the main analysis). Previous analyses showing extensive atrophy in the occipital lobe failed to control for CAG length (Tabrizi et al. 2009; Tabrizi et al. 2012) and its inclusion here suggests that the occipital lobe may be more susceptible to the negative effects of high CAG than other regions.

Finally, the behavioural analysis links change measured within the brain to deterioration on motor and cognitive scores. TMS is a clinical measure of motor performance, that covers a range of motor domains representing clinical progression in HD, including chorea, dystonia, ocular pursuit, and gait. Here, TMS showed accelerated impairment prior to motor onset suggesting that performance on this measure, and thus typical HD movement symptoms are more closely linked to striatal atrophy than that within the cortex, and more likely the result of reduced connectivity as demonstrated by McColgan et al. (2017). In contrast, both the Speeded Tapping and SDMT measures showed an accelerated loss in performance mainly after clinical diagnosis. Both measures, designed to quantify more discrete aspects of motor and cognitive performance than TMS, therefore, begin to show greater impairment as cortical motor regions show accelerated volume loss, thus indicating that performance on both the Speeded Tapping and SDMT could be more closely linked to cortical atrophy than TMS.

This work has taken a step towards understanding the progression of cortical atrophy in HD and has implications for the treatment of HD. The results suggest that cortical atrophy is undergoing significant change in the 5 years preceding clinical diagnosis, and thus support the view that administration of a potential therapy aimed at preventing degeneration should be timed many years prior to onset. In addition, preclinical work conducted on primates suggests that, if successful, the intrathecal administration of *HTT* lowering drugs currently being trialled on early HD participants will have a greater effect on reducing *HTT* in the cortex than the striatum (Wild & Tabrizi 2017). Consequently, characterising the pattern of cortical atrophy in HD is essential for understanding the effectiveness of these therapies at slowing or preventing atrophy. As research moves towards the development of a successful disease-modifying treatment for HD, this knowledge may prove useful for the refinement of the timing and administration of a treatment, especially if this type of analyses is conducted on other HD cohorts at different disease stages. Although the relationship between neural atrophy and the behavioural measures studied here should ultimately be tested more directly, these findings have important implications for understanding the association between neural changes and commonly used HD biomarkers for measuring symptom progression. Despite its sensitivity to early motor symptoms (Tabrizi et al. 2012), TMS is routinely used to measure change in groups of HD participants rather than pre-HD participants, and as such is used as an endpoint in clinical trials. However, as these findings demonstrate, TMS begins to show increased change prior to any significant acceleration in cortical atrophy, and therefore, it is possible that it may not be as sensitive to the effects of a drug that targets the cortex. In contrast, changes in Speeded Tapping and SDMT suggest a closer association with changes in cortical atrophy of

the motor regions, suggesting that they are more sensitive to the effects of the therapies currently being tested, and this should, therefore, be considered during analysis of trial endpoints.

This study has improved on previous work by addressing a number of limitations observed in other studies. The modelling framework applied here is a previously validated, yet novel approach to modelling sMRI data and this study is the first application of this technique to neurodegeneration. The method was developed to address weaknesses in previous analysis methods and to approach the quantification of GM change via a dynamical systems method. The ability to quantify both linear and non-linear CGM change in a group of participants with data from variable visits over multiple time points offers a more powerful approach to CGM modelling than previous methods. Using this approach, the quantification of linear atrophy as well as acceleration of atrophy results in a detailed perspective on the progression of neural change than previous methodologies applied in HD. Different generative models were fitted to the data and compared, so that traditional models such as a quadratic model fit could be compared with more complex models, including the sigmoid model. The sigmoidal model was found to have the best data fit once incorporating varying delay and acceleration factors. By performing Bayesian model comparison on the different models, it is ensured that the results of this comparison are robust. Furthermore, the inclusion of covariates into the first-level of this model controls for variability in these parameters, and thus the models are less affected by individual differences. While it is essential that this work is replicated using similar approaches in other datasets and cohorts, the regional progression observed in these results is entirely plausible based on post-mortem data, previous imaging data and disease symptoms.

By performing this analysis on a subset of TRACK-HD and TrackOn-HD participants who underwent conversion, some heterogeneity within the disease progression can be removed. All participants in this study were showing clinical progression over this period, thus removing the variability introduced by using a predictor of motor onset to categorize participants. The final dataset included 286 scans across 49 participants, resulting in over a decade of data surrounding motor diagnosis. This timespan and number of participants provides a strong dataset for quantifying longitudinal within-participant change. PREDICT-HD investigators collected data on over 1000 pre-HD participants for >10 years, and are the only other known cohort with such a large extent of imaging data in HD. One study conducted on the PREDICT-HD dataset examined motor, cognitive and imaging predictors of motor diagnosis (Paulsen, Long, Ross, et al. 2014). Variables quantified at baseline from all three domains were significant predictors of conversion from pre- to manifest HD in 225 individuals. A number of

brain regions were significant predictors of conversion (including the putamen, caudate, hippocampus), but lobular GM was one of the best predictors. However, lobular regions were not broken down further, and there has not yet been any analysis that can compare to the current longitudinal analysis of cortical sub-regions, however the PREDICT-HD dataset offers the potential to replicate this result in a much larger cohort of motor converters. While the current analysis shows some interesting results, they require further validation in different cohorts and across different stages of HD before generalising them.

Finally, the application of robust segmentation and analysis methods add support to the current findings. By first validating MALP-EM on the TRACK-HD cohort and performing detailed visual QC of all scans, the quality of segmentations has been assured. Since it is clear that substandard segmentation methods have previously lead to mixed findings in neuroimaging studies (Ashburner et al. 2016), the use of accurate tools that were validated on the same cohort in which they were then applied is an important strength of this study.

It is also important to address a number of weaknesses in this study. Firstly, the use of clinical motor diagnosis as a criterion for inclusion in this study provides a clinically meaningful measure of disease progression, but also introduces a level of subjectivity to the inclusion of participants. The Leiden site had many more converters compared to other sites, with participants from Leiden almost 45% of participants in this analysis. It is unclear whether this is due to differences in clinical diagnosis procedures between raters or countries or whether more participants did undergo change in Leiden. In order to control for this where possible, site was entered as a co-variate into the analyses. In addition, the inclusion of participants within 6 years to motor onset means that it is likely that the very earliest cortical atrophy has not been captured by the current analysis. Further analyses could attempt to use pre-HD participants to model cortical atrophy in those further from onset, although this re-introduces issues associated with using predicted onset as a group classifier. Finally, this analysis combined data from two hemispheres, and merged a number of smaller cortical regions into larger regions. These decisions were made prior to analysis to reduce the number of regions included in the analysis, and were based on previous evidence suggesting that CGM changes in HD are bilateral. Some regions that were combined may more effectively represent HD without being combined, especially the cingulate gyrus. Analysis should be repeated with different parcellations to determine how this affects the results found in the current study. Nevertheless, it is possible that subtle hemispheric differences or small regional differences were overlooked due to this decision.

This analysis has given new insight into the progression of cortical atrophy, yet there are a number of future aims that could be addressed to build on these findings and address weaknesses in this study. As discussed, the analysis could be replicated within smaller sub-regions of the cortex and bilaterally to determine how this might affect the results. In addition, future work can build on the current examination by quantifying change within different cohorts. For example, a recently developed score of HD progression that was validated on the TRACK-HD cohort could be used to select those who undergo fast or slow HD progression (Moss et al. 2017). In addition, while there appear to be distinct HD phenotypes that occur around the time of conversion that would be interesting to study (McCusker & Loy 2017), it can be very challenging to group participants based on symptom progression, especially in a relatively small sample size. The present analysis was limited to the analysis of sMRI data with a focus on the cortex. Future analyses could work to integrate other subcortical regions and imaging metrics or modalities such as measures of diffusion or connectivity into the model. This would provide a more complete understanding of the process of neurological change occurring in HD, rather than just cortical atrophy. Finally, the behavioural analysis performed in this chapter should be combined with the volumetric sDCM modelling. Directed hypotheses could be tested based on the preliminary results indicating that TMS may show early changes related to WM (i.e. structural connectivity) and striatal changes, but that Speeded Tapping and the SDMT may be more closely associated with volumetric loss in the cortex.

These results provide the most detailed characterization of cortical atrophy in participants undergoing transition from pre- to manifest HD to date. By applying a recently validated model that is uniquely able to map both linear and non-linear cortical change across multiple time points, this analysis expands upon previous work quantifying atrophy within the cortex. The results suggest that both linear and non-linear changes contribute to cortical atrophy differentially across the brain, and provide preliminary results indicating that different behavioural measures are likely to show different associations with cortical atrophy. While future work could expand upon these results within different cohorts and by integrating different imaging measures, these results presenting the first brain-wide cortical mapping in HD provide a comprehensive insight into CGM atrophy at a critical point in HD progression.

7. DISCUSSION

This thesis aimed to evaluate the most accurate tools for measuring cortical atrophy from MRI scans in HD and to then apply the validated tools on an HD cohort undergoing motor conversion in an attempt to characterise cortical grey matter (CGM) change during this period. First, a series of comparisons between commonly used CGM and subcortical GM measurement tools were performed. From this, the most appropriate methods to detect and measure GM change in HD were selected. These tools were then applied to data from a cohort of HD gene carriers who underwent motor conversion during the TRACK-HD and TrackOn-HD studies. The data were analysed using a newly developed structural dynamic causal modelling (DCM) modelling technique that provided a detailed account of what CGM changes occur during the period immediately surrounding HD diagnosis. The results from this thesis provide both methodological advances in the analysis of sMRI data and clinically relevant information about the progression of atrophy in HD.

7.1. Methodological comparisons

Three methodological comparisons were performed in chapters three, four and five. Chapter three compared a number of methods that can be used to quantify the volume of the CGM. The measurement of brain volume is the most commonly used method of quantifying neural atrophy in HD, particularly within subcortical regions of interest (ROIs) (Georgiou-Karistianis et al. 2013). Previous work has demonstrated that variability in structural brain volume measures introduced by the application of different image analysis tools can have large effects on the conclusions drawn in neuroimaging studies (Katuwal et al. 2016) and thus the use of accurate measures are paramount in quantifying cortical volume change in HD. Seven segmentation tools were compared using two datasets; an artificial dataset with corresponding 'ground truth' (GT) and a subset of TRACK-HD data that included control, pre-HD and manifest-HD participants. Thorough qualitative and quantitative comparisons were undertaken and the results provided specific evidence for the selection of a segmentation tool for use in this thesis as well as a series of more widely applicable recommendations for other researchers who are also trying to select the most appropriate method of segmentation for their data. The comparison demonstrated that while all tools performed reliably across multiple scans, accuracy was much more variable. During visual QC it was clear that all tools had examples of poor segmentation performance, with some tools generally performing very poorly. This is particularly concerning given that some of the most widely used tools

(FreeSurfer, FSL FAST, SPM 8 Unified Segment) showed consistent errors in segmentation that are likely to have a considerable effect on the results of any study using these techniques. This finding was further established by a quantitative analysis that supported previous work in showing that the results of statistical comparisons differed between techniques, and thus that the application of different tools to measure CGM could result in varying conclusions when studying disease progression in HD (Katuwal et al. 2016). For example, when compared to controls there was significant longitudinal change in the CGM for pre-HD TRACK-HD participants, which was detected by the segmentations from by some tools but not others. This result stems from the differences in volumetric regions that are generated from different segmentation techniques and highlights the importance of choosing the best segmentation tool for the data, and then performing visual QC to ensure that segmentations are accurate before drawing conclusions based on the regions.

The tool that demonstrated the highest standard of segmentation, both quantitatively and qualitatively on both datasets was MALP-EM. It is a newly developed, easy-to-use tool that was validated for use on traumatic brain injury (Ledig et al. 2015). MALP-EM is a multi-atlas based technique that is suitable for use on clinical populations, as demonstrated by the previous successful application in patients suffering from severe examples of traumatic brain injury (Ledig et al. 2015). It had high accuracy when quantified via comparing the segmentations to the BrainWeb GT segmentations, high reliability as demonstrated by test-retest TRACK-HD data, and was sensitive to volumetric change in the longitudinal TRACK-HD analysis. Finally, the segmentations were generally better than most other techniques based on visual QC. While MALP-EM did show some errors, particularly in the temporal and occipital regions, they were much more infrequent and minor than for other tools and can easily be improved via the inclusion of a brain mask in the segmentation pipeline if needed. Based on the results from this chapter MALP-EM was selected as the segmentation tool that would be used to measure CGM in this thesis. It was the most visually accurate measurement tool, but also highly reliable and easy to apply, which will hopefully facilitate easy replication in other datasets.

In chapter four another methodological comparison was undertaken to examine the performance of measures of cortical thickness (CT). It has been hypothesised that CT may be a more sensitive measure of cortical change than GM since volumetric measures are a function of the relationship between the surface area of the brain and CT (Panizzon et al. 2009), yet CT is a difficult characteristic to accurately quantify. Due to the complex nature of the cortex, CT is often over-estimated. This is largely because it is challenging to accurately delineate the tightly

bound sulci of a brain. Both surface-based and voxel-based methods of quantifying CT are available, with surface-based methods considered to be more accurate but time-consuming to process. This chapter compared three measures of CT, one surface-based method and two voxel-based methods. While the surface-based tool, FreeSurfer, gave promising longitudinal results that indicated the occurrence of cortical thinning with increasing disease progression in HD, the segmentations were substandard when visually QCed. There were regular errors in the classification of WM and GM, particularly in the occipital lobe, and an overall trend for the underestimation of the GM. Of the two voxel-wise measures tested, one was highly reliable but showed very poor longitudinal performance (MALP-EM segmentation + ANTs CT measurement), and the other was generally unreliable for both cross-sectional and longitudinal measurement of CT (ANTs). These disappointing results highlight the need for further development of accurate CT measures, and suggest that when using the currently available techniques caution should be taken. While FreeSurfer may be able to successfully distinguish between different groups and measure changes in CT, as was demonstrated here and in previous studies (Righart et al. 2017), the issues associated with the quality of segmentations should be considered, especially when comparing a group with neural pathology to a healthy control group. The visual examination of all segmentations should be performed and the results of a study interpreted with these in mind. While it is possible to perform manual edits within FreeSurfer that may improve the segmentations, as research studies expand the number of participants recruited and the move to analysing large cross-study datasets becomes more popular this is an unfeasible solution. While it was initially hoped that CT could be used as an additional and potentially more sensitive longitudinal measure for detecting cortical change in HD in this thesis, it was instead excluded from subsequent chapters based on the results of chapter four.

The final methodological comparison was performed in chapter five, whereby manually delineated volumetric regions of the caudate and putamen were compared to the automated segmentations produced by MALP-EM. Manually delineated subcortical regions are often considered a ‘gold standard’, but are time-consuming to perform (both in terms of rater training and completing the measurements). Automated measures are regularly used within HD literature due to their accessibility and speed of use, but the MALP-EM subcortical regions have not been used in HD work previously and they have not been compared to manual measures. After running MALP-EM on a subset of the data from PADDINGTON, visual QC indicated that MALP-EM subcortical regions were generally of high quality. There were a small number of failed segmentations, however these could be improved via manual edits to the

regions. The quantitative analysis demonstrated good overlap between the manual and automatically generated regions, with a very close association between the volumes from both tools. The results from this chapter provided validation of the MALP-EM regions for use in this study.

These three analyses were performed to ensure that the investigation undertaken in chapter six used highly accurate measures of CGM to characterise GM change. The accurate measurement of neural characteristics is of utmost importance when performing sMRI studies, especially when researching clinical cohorts whereby conclusions may influence the understanding of disease pathology. Previous work has demonstrated that the use of different techniques to measure brain volume could be driving variability in neuroimaging findings (Katuwal et al. 2016), and encouraged the cross-validation and consistency when using segmentation techniques to measure volume. While numerous methodological comparisons have been performed to validate both volumetric and CT methods (Clarkson et al. 2011; Schwarz et al. 2016; Fellhauer et al. 2015; Kazemi & Noorizadeh 2014; Despotović et al. 2015; Klauschen et al. 2009; Eggert et al. 2012), the work undertaken for this thesis also placed importance on the visual examination of segmentation quality. The qualitative review of segmentations has previously been overlooked during reporting on the performance of many tools, with few published studies providing descriptions of the qualitative review of segmentation accuracy. Here, the qualitative examination of >3000 segmentations resulted in a good understanding of the limits of currently available automated tools. The majority of tools, both for the measurement of volume and CT, had issues with delineating the occipital and temporal lobes with over-estimation of CGM common within these regions. Despite many available software options, automated techniques are not always able to delineate the complex structure of the human brain. Understanding the weaknesses of these tools is vital when using them, and other users should endeavour to perform careful visual examination of both volumetric and CT segmentations prior to publishing studies that use the tools.

The work in this thesis also identified the strengths of volumetric measurements compared to CT measurements, with findings suggesting that while, theoretically, measures of volume may be less sensitive to neural change than measures of CT (Panizzon et al. 2009), volume is easier to quantify and more robust to errors in segmentation than current CT measures. CT measures, especially when quantified longitudinally, are particularly susceptible to the influence of errors in segmentation that inflate CT values. Errors include the poor delineation of tightly bound sulci as well as general over- or under-estimation of the cortex due to noise within scans. While similar findings have been discussed previously (Clarkson et

al. 2011; Righart et al. 2017), the current results suggest that the use of recently developed CT methods and highly validated cortical regions do not address the shortfalls previously seen in CT studies. Until more work is done to improve measures of CT, volumetric measures may be more appropriate for the detection of cortical atrophy.

7.2. Mapping cortical change during HD progression

The final chapter of this thesis applied the methods validated in chapters three and five to a subset of data from the TRACK-HD and TrackOn-HD studies. Atrophy of the cortex has not been extensively characterised in HD, with previous work suffering from a range of limitations. This study aimed to expand upon past work and present a detailed characterisation of cortical atrophy during the transition from pre-HD to manifest HD, a phase where cortical atrophy is thought to increase along with increases in HD symptoms. Scans from 49 participants with multiple time points were segmented, and CGM change during the period prior to and after HD motor diagnosis was quantified. A dynamical model using a DCM framework was applied that quantified total CGM loss during the period, along with the rate of linear (constant) atrophy, the rate and timing of non-linear atrophy (accelerations of atrophy) and the relationship between atrophy and CAG across the cortex. This combination of results provides the first in-depth understanding of the complexity of regional atrophy in the cortex and how this changes over time, particularly around the critical time of symptom onset in HD.

The results show that volume loss in various regions of the cortex is higher than previously thought. A number of frontal and motor regions underwent atrophy totalling around 10% during the decade around motor diagnosis. This level of loss is double that at which cortical atrophy has previously been estimated (Tabrizi et al. 2012), indicating that heterogeneity of atrophy within the cortex influences global rates of cortical atrophy. When regional rates of linear atrophy were quantified, regions of the frontal lobe were shown to undergo the highest rates of linear atrophy in the cortex. Neuronal loss in the frontal lobe has been associated with many of the earliest symptoms of HD (Nana et al. 2014), and reduced volume of the frontal lobe is associated with poorer performance on a number of cognitive and motor tasks in HD (Scahill et al. 2013). However, the finding that frontal regions are undergoing the highest rate of linear atrophy in this vital period of HD progression has not been demonstrated before. In addition, motor regions appear to undergo the greatest acceleration of atrophy during transition from pre-HD to manifest HD. The clinical categorisation of HD is based upon increasing motor symptoms and it is thus perhaps unsurprising that these regions undergo accelerations during this period, yet the extent of acceleration in these regions compared to other regions is very striking. In contrast, subcortical

regions were undergoing consistently high rates of atrophy during the whole period, with very little acceleration present, supporting previous evidence that has suggested low rates of acceleration of atrophy in the caudate (Hobbs, Barnes, et al. 2010). The timing of the accelerated atrophy in motor regions appears to occur slightly after motor onset, indicating that the cause of symptom onset is not CGM atrophy but possibly the breakdown in WM connections, as suggested by previous work (Wu et al. 2017; McColgan et al. 2017). Together, these two results give an interesting insight into the neural changes underpinning HD progression.

The association between cortical progression and CAG length was also modelled. The impact of CAG on neural atrophy has not been extensively characterised, despite evidence that CAG length influences both HD onset and progression (Ross et al. 2014). Previous work has used VBM and linear models to suggest that there is a relationship between atrophy and CAG length, especially in subcortical regions, with some evidence for associations between CAG length and atrophy in the occipital lobe. The results confirm that higher CAG length is associated with faster progression in the occipital lobe in this cohort. Furthermore, this analysis also showed that those with higher CAG lengths underwent less acceleration of atrophy. That is, they underwent a more constant rate of atrophy compared to those who had lower CAG lengths and showed an acceleration of atrophy. This result is new evidence that indicates that there may be a different pattern of atrophy in those with high vs low CAG repeat lengths.

Finally, a behavioural analysis offers preliminary results linking cortical atrophy to three commonly used measures of HD progression, the Total Motor Score (TMS), Speeded Tapping, and SDMT. While the results require further analysis using a more directed model, they provide an indication that Speeded Tapping and SDMT, discrete measures of HD progression, are more closely linked to cortical atrophy especially within the motor regions than TMS, which undergoes earlier change in this cohort, indicative of a possible link to WM connectivity changes described previously.

The results from this chapter give new insight into the complex nature of regional CGM atrophy occurring over disease progression. By using a highly optimised segmentation technique with a complex modelling method to expand upon previous work that was conducted using linear modelling and more heterogeneous groups this analysis delivers a novel approach. The results demonstrate the importance of both frontal and motor regions in the progression of HD. Although it is possible that the atrophy within these regions is not causing the onset of symptoms, it may hasten their progression. The behavioural analysis

warrants further investigation, but suggests that discrete measures of HD progression may be more closely linked to cortical change than measures quantifying change across multiple domains. That there are significant accelerations of atrophy in sub-regions of the cortex during this phase of disease progression indicates that that CGM atrophy is not linear, as previously assumed and shown in Figure 1.6 from Chapter One, and that this is an oversimplification. The therapeutic trials that are currently underway are likely to target the cortex, and thus by having a better understanding of the nature of cortical atrophy and which regions are showing the greatest progression, as well as associated change on behavioural biomarkers, a more targeted approach to quantifying the effects of the drug could be used.

7.3. Weaknesses and future work

It is important to recognise a number of weaknesses in this work, these are covered in detail in the preceding chapters, but a few key points are reiterated here. The methodological comparisons did not compare longitudinal pipelines, since these were not available for all methods. By using longitudinal pipelines the results from some techniques could be improved and this should be considered when selecting methods for analysis. Furthermore, the methodological validations performed in chapters three and four did not include the comparison of automated methods to a gold standard for CGM regions. While a manual segmentation technique was tested, it was not found to be a reproducible or reliable measure. Future work should continue to develop improved techniques and compare sMRI methodology. The work conducted in this thesis has demonstrated that the techniques used to measure neural characteristics require improvement, particularly as MRI acquisitions develop. The increasing access to 7T data is just one example of a rapidly changing field, and it is essential that the tools used to analyse this data are constantly evaluated and improved.

The TRACK-HD, TrackOn-HD and PADDINGTON studies were multi-site studies. While site was controlled for in analyses, there are a number of factors (scanner drift, software upgrades) that will not be accounted for by using a dummy variable to control for site. The acquisitions were developed for multi-site use and regular quality assurance was performed on each scanner to maintain consistency within and between sites, however it is likely that some issues are introduced by site differences.

Replication of the sDCM work conducted in chapter six is important. Ideally, this work would be replicated in a different cohort, such as the PREDICT-HD cohort. The addition of other imaging data such as connectivity metrics could further enhance the understanding of disease progression in this cohort, as would further modelling to directly link behavioural change to neural change. By using one model to map the relationship between different

measures of neural change and how they change over time a cohesive picture of the progression of brain changes in HD can be developed. In addition, the use of different cohorts could provide further insight on the findings shown here. The comparison of a high CAG vs low CAG group could offer further evidence of the relationship between CAG and atrophy. In addition, the relationship between CGM atrophy and change in clinical and cognitive variables should be more thoroughly examined.

7.4. Conclusions

This thesis has provided methodological advances on the quantification of neural characteristics from sMRI data in HD, with the resulting methods then applied on a clinical cohort to conduct a thorough characterisation of the trajectory of CGM atrophy during HD motor symptom onset. The results provide important validation of a recently developed and easy to use segmentation method, and present important recommendations for the application of a number of commonly used sMRI measurement techniques. Most importantly, this thesis has advanced the current understanding of CGM atrophy in HD, revealing that rates of cortical atrophy are higher than previously thought. The pattern of atrophy seen across the cortex reveals the complex nature of cortical change occurring during motor conversion in HD. The results have implications for the overall understanding of disease progression, but also offer potential value in the monitoring of cortical change during clinical trials aimed at developing a treatment for HD.

8. REFERENCES

Alfaro-Almagro, F. et al., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166, pp.400–424.

Anon, 1994. Guidelines for the molecular genetics predictive test in Huntington's disease. International Huntington Association (IHA) and the World Federation of Neurology (WFN) Research Group on Huntington's Chorea. *Neurology*, 44(8), pp.1533–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8058167 [Accessed October 22, 2017].

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1), pp.95–113. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17761438 [Accessed August 3, 2017].

Ashburner, J. et al., 2016. Inter-Method Discrepancies in Brain Volume Estimation May Drive Inconsistent Findings in Autism. *Frontiers in Neuroscience Front. Neurosci*, 10(10).

Ashburner, J. & Friston, K.J., 2005. Unified segmentation. *NeuroImage*, 26(3), pp.839–51. Available at: http://www.sciencedirect.com/science/article/pii/S1053811905001102 [Accessed July 11, 2014].

Ashburner, J. & Friston, K.J., 2000. Voxel-based morphometry--the methods. *NeuroImage.*, 11, pp.805–821.

Ashburner, J. & Ridgway, G.R., 2012. Symmetric diffeomorphic modeling of longitudinal structural MRI. *Frontiers in neuroscience*, 6, p.197. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23386806 [Accessed October 24, 2016].

Aubert-Broche, B. et al., 2006. Twenty new digital brain phantoms for creation of validation image data bases. *IEEE transactions on medical imaging*, 25(11), pp.1410–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17117770 [Accessed February 13, 2015].

Avants, B.B., Tustison, N.J., Song, G., et al., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3), pp.2033–2044.

Avants, B.B., Tustison, N.J., Wu, J., et al., 2011. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*, 9(4), pp.381–400. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21373993 [Accessed August 30, 2016].

Aylward, E.H., Mills, J., et al., 2011. Association between age and striatal volume Stratified by CAG repeat length in prodromal Huntington disease. *PLoS Currents*.

Aylward, E.H. et al., 1998. Frontal lobe volume in patients with Huntington's disease. *Neurology*, 50(1), pp.252–258. Available at: http://www.neurology.org/cgi/doi/10.1212/WNL.50.1.252 [Accessed May 8, 2017].

Aylward, E.H. et al., 1997. Longitudinal change in basal ganglia volume in patients with Huntington's disease . *Neurology* , 48 (2), pp.394–399.

Aylward, E.H., Nopoulos, P.C., et al., 2011. Longitudinal change in regional brain volumes in prodromal Huntington disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 82(4), pp.405–410.

Aylward, E.H. et al., 2004. Onset and rate of striatal atrophy in preclinical Huntington disease.

Neurology, 63, pp.66–72.

Aylward, E.H. et al., 2012. Striatal volume contributes to the prediction of onset of Huntington disease in incident cases. *Biol.Psychiatry*, 71, pp.822–828.

Barnes, J. et al., 2010. Head size, age and gender adjustment in MRI studies: a necessary nuisance? *NeuroImage*, 53(4), pp.1244–1255. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20600995 [Accessed September 26, 2017].

Bartko, J.J., 1966. The Intraclass Correlation Coefficient As a Measure of Reliability. *Psychological Reports*, 19(1), pp.3–11.

Bates, G.P. et al., 2015. Huntington disease. *Nature Reviews Disease Primers*, p.15005. Available at: http://www.nature.com/articles/nrdp20155 [Accessed October 2, 2015].

Beeri, M.S. et al., 2006. Age, gender, and education norms on the CERAD neuropsychological battery in the oldest old. *Neurology*, 67(6), pp.1006–1010.

Le Bihan, D. et al., 2001. Diffusion tensor imaging: concepts and applications. *Journal of magnetic resonance imaging : JMRI*, 13(4), pp.534–46. Available at: http://europepmc.org/abstract/med/11276097 [Accessed November 18, 2015].

Björkqvist, M. et al., 2008. A novel pathogenic pathway of immune activation detectable before clinical onset in Huntington's disease. *The Journal of experimental medicine*, 205(8), pp.1869–77. Available at: http://jem.rupress.org/content/205/8/1869.short [Accessed January 19, 2016].

Bouix, S. et al., 2007. On evaluating brain tissue classifiers without a ground truth. *NeuroImage*, 36(4), pp.1207–1224. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17532646 [Accessed September 26, 2017].

Boyes, R.G. et al., 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. *NeuroImage*, 39(4), pp.1752–62. Available at: http://www.sciencedirect.com/science/article/pii/S1053811907009494 [Accessed December 1, 2014].

Buckner, R.L. et al., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(5), pp.2322–45. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3214121&tool=pmcentrez&rendertype=abstract [Accessed July 23, 2015].

Byrne, L.M. et al., 2017. Neurofilament light protein in blood as a potential biomarker of neurodegeneration in Huntington's disease: a retrospective cohort analysis. *The Lancet Neurology*, 16(8), pp.601–609. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28601473 [Accessed August 3, 2017].

Byrne, L.M. & Wild, E.J., 2016. Cerebrospinal Fluid Biomarkers for Huntington's Disease. *Journal of Huntington's Disease*, 5(1), pp.1–13. Available at: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JHD-160196 [Accessed November 28, 2017].

Callaert, D. V. et al., 2014. Assessing age-related gray matter decline with voxel-based morphometry depends significantly on segmentation and normalization procedures. *Frontiers in Aging Neuroscience*, 6, p.124. Available at: http://journal.frontiersin.org/article/10.3389/fnagi.2014.00124/abstract [Accessed October 12, 2016].

Carroll, J.B. et al., 2015. Treating the whole body in Huntington's disease. *The Lancet.*

Neurology, 14(11), pp.1135–42. Available at: http://www.sciencedirect.com/science/article/pii/S1474442215001775 [Accessed October 16, 2015].

Ciarochi, J.A. et al., 2016. Patterns of Co-Occurring Gray Matter Concentration Loss across the Huntington Disease Prodrome. *Frontiers in neurology*, 7, p.147. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27708610 [Accessed October 15, 2017].

Clarkson, M.J. et al., 2011. A comparison of voxel and surface based cortical thickness estimation methods. *NeuroImage*, 57(3), pp.856–865. Available at: http://www.sciencedirect.com/science/article/pii/S1053811911005611 [Accessed May 8, 2017].

Coppen, E.M. et al., 2016. Early grey matter changes in structural covariance networks in Huntington's disease. *NeuroImage: Clinical*, 12, pp.806–814.

Crum, W.R., Camara, O. & Hill, D.L.G., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, 25(11), pp.1451–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17117774 [Accessed December 12, 2015].

Crum, W.R., Scahill, R.I. & Fox, N.C., 2001. Automated Hippocampal Segmentation by Regional Fluid Registration of Serial MRI: Validation and Application in Alzheimer's Disease. *NeuroImage*, 13(5), pp.847–855. Available at: http://www.sciencedirect.com/science/article/pii/S105381190190744X [Accessed January 9, 2018].

Currie, S. et al., 2013. Understanding MRI: basic MR physics for physicians. *Postgraduate Medical Journal*, 89(1050), pp.209–223. Available at: http://pmj.bmj.com/lookup/doi/10.1136/postgradmedj-2012-131342 [Accessed November 13, 2017].

Dale, A.M., Fischl, B. & Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9, pp.179–194.

Das, S.R. et al., 2009. Registration based cortical thickness measurement. *NeuroImage*, 45(3), pp.867–879. Available at: http://www.sciencedirect.com/science/article/pii/S1053811908012780 [Accessed January 20, 2016].

Derakhshan, M. et al., 2010. Evaluation of automated techniques for the quantification of grey matter atrophy in patients with multiple sclerosis. *NeuroImage*, 52(4), pp.1261–1267.

Desikan, R.S. et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3), pp.968–980. Available at: http://www.sciencedirect.com/science/article/pii/S1053811906000437 [Accessed May 8, 2017].

Despotović, I., Goossens, B. & Philips, W., 2015. MRI segmentation of the human brain: Challenges, methods, and applications. *Computational and Mathematical Methods in Medicine*, 2015.

Dogan, I. et al., 2013. Consistent Neurodegeneration and Its Association with Clinical Progression in Huntington's Disease: A Coordinate-Based Meta-Analysis. *Neurodegenerative Diseases*, 12(1), pp.23–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22922585 [Accessed November 29, 2017].

Domínguez, J.F. et al., 2013. Multi-Modal Neuroimaging in Premanifest and Early Huntington's

Disease: 18 Month Longitudinal Data from the IMAGE-HD Study B. Draganski, ed. *PLoS ONE*, 8(9), p.e74131. Available at: http://dx.plos.org/10.1371/journal.pone.0074131 [Accessed September 28, 2016].

Douaud, G. et al., 2006. Distribution of grey matter atrophy in Huntington's disease patients: A combined ROI-based and voxel-based morphometric study. *NeuroImage*, 32(4), pp.1562–1575. Available at: http://www.sciencedirect.com/science/article/pii/S1053811906006422 [Accessed November 29, 2017].

Eggert, L.D. et al., 2012. Accuracy and reliability of automated gray matter segmentation pathways on real and simulated structural magnetic resonance images of the human brain. *PLoS One*, 7(9), p.e45081. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23028771 [Accessed August 30, 2016].

Evans, S.J. et al., 2013. Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. *Journal of Neurology, Neurosurgery & Psychiatry*, 84(10), pp.1156–1160. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23482661 [Accessed October 17, 2017].

Feigin, A. et al., 2001. Metabolic network abnormalities in early Huntington's disease: an [(18)F]FDG PET study. *J.Nucl.Med.*, 42, pp.1591–1595.

Fellhauer, I. et al., 2015. Comparison of automated brain segmentation using a brain phantom and patients with early Alzheimer's dementia or mild cognitive impairment. , 233(3), pp.299–305.

Fischl, B. et al., 2004. Automatically parcellating the human cerebral cortex. *Cereb Cortex*, 14, pp.11–22.

Fischl, B. et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3), pp.341–355. Available at: http://www.sciencedirect.com/science/article/pii/S089662730200569X [Accessed May 8, 2017].

Fischl, B. & Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc.Natl.Acad.Sci.U.S.A*, 97, pp.11050–11055.

Fischl, B., Sereno, M.I. & Dale, A.M., 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9, pp.195–207.

Fisher, E.R. & Hayden, M.R., 2014. Multisource ascertainment of Huntington disease in Canada: Prevalence and population at risk. *Movement Disorders*, 29(1), pp.105–114. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24151181 [Accessed October 17, 2017].

Fjell, A.M. et al., 2015. Development and aging of cortical thickness correspond to genetic organization patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50), pp.15462–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26575625 [Accessed April 7, 2017].

Focke, N.K. et al., 2011. Multi-site voxel-based morphometry — Not quite there yet. *NeuroImage*, 56, pp.1164–1170. Available at: https://ac.els-cdn.com/S1053811911001856/1-s2.0-S1053811911001856-main.pdf?_tid=e97baa66-f776-11e7-8f06-00000aab0f6c&acdnat=1515747764_7a6f868dd6f60f3388c96fc88a988a35 [Accessed January 12, 2018].

Freeborough, P.A. & Fox, N.C., 1998. Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images. *Journal of computer assisted tomography*, 22(5), pp.838–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9754126 [Accessed August 3, 2017].

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, pp.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. *Comput.Methods Programs Biomed.*, 53, pp.15–25. Available at: file://m/ION/TrackHD/Imaging/Eli/database/Interactive_alogorithms_for_the_segmentation.pdf.

Friston, K.J. et al., 2016. Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage*, 128, pp.413–431.

Friston, K.J. et al., 2007. Variational free energy and the Laplace approximation. *NeuroImage*, 34(1), pp.220–234.

Friston, K.J., Harrison, L. & Penny, W., 2003. Dynamic causal modelling. *NeuroImage*, 19(4), pp.1273–1302.

Gavazzi, C. et al., 2007. Combining Functional and Structural Brain Magnetic Resonance Imaging in Huntington Disease. *Journal of Computer Assisted Tomography*, 31(4), pp.574–580. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17882035 [Accessed November 30, 2017].

Georgiou-Karistianis, N. et al., 2013. Structural MRI in Huntington's disease and recommendations for its potential use in clinical trials. *Neuroscience and Biobehavioral Reviews*, 37(3), pp.480–490. Available at: http://dx.doi.org/10.1016/j.neubiorev.2013.01.022.

Gómez-Ansón, B. et al., 2009. Prefrontal cortex volume reduction on MRI in preclinical Huntington's disease relates to visuomotor performance and CAG number. *Parkinsonism and Related Disorders*, 15(3), pp.213–219.

Gregory, S. et al., 2017. Operationalizing compensation over time in neurodegenerative disease. *Brain*, 140(4), pp.1158–1165.

Gronenschild, E.H.B.M. et al., 2012. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements S. Hayasaka, ed. *PLoS One*, 7(6), p.e38234. Available at: http://dx.plos.org/10.1371/journal.pone.0038234 [Accessed May 8, 2017].

Habeck, C.G., 2010. Basics of Multivariate Analysis in Neuroimaging Data. *Journal of Visualized Experiments*, (41). Available at: http://www.ncbi.nlm.nih.gov/pubmed/20689509 [Accessed November 30, 2017].

Halliday, G. et al., 1998. Regional Specificity of Brain Atrophy in Huntington's Disease. *Experimental Neurology*, 154(2), pp.663–672. Available at: http://www.sciencedirect.com/science/article/pii/S0014488698969199 [Accessed November 21, 2017].

Han, X. et al., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32, pp.180–194.

Harrington, D.L. et al., 2014. Neuroanatomical correlates of cognitive functioning in prodromal Huntington disease . *Brain and Behaviour* , 4 (1), pp.29–40.

Heckemann, R.A. et al., 2012. Multi-atlas propagation with enhanced registration { MAPER. *MICCAI 2012 Workshop on Multi-Atlas Labeling*, p.83. Available at: https://hal.archives-ouvertes.fr/hal-00839257/ [Accessed January 22, 2016].

Henley, S.M. et al., 2006. Increased rate of whole-brain atrophy over 6 months in early Huntington disease. *Neurology*, 67, pp.694–696.

Henley, S.M. et al., 2010. Pitfalls in the use of voxel-based morphometry as a biomarker: examples from huntington disease. *AJNR Am.J Neuroradiol.*, 31, pp.711–719.

Henley, S.M. et al., 2009. Relationship between CAG repeat length and brain volume in premanifest and early Huntington's disease. *J Neurol*, 256, pp.203–212.

Ho, A.K. et al., 2009. Health-related quality of life in Huntington's disease: Which factors matter most? *Movement Disorders*, 24(4), pp.574–578.

Hobbs, N.Z. et al., 2009. Automated quantification of caudate atrophy by local registration of serial MRI: evaluation and application in Huntington's disease. *NeuroImage*, 47, pp.1659–1665.

Hobbs, N.Z. et al., 2013. Evaluation of multi-modal, multi-site neuroimaging measures in Huntington's disease: Baseline results from the PADDINGTON study. *NeuroImage: Clinical*, 2, pp.204–211.

Hobbs, N.Z., Barnes, J., et al., 2010. Onset and progression of pathologic atrophy in Huntington disease: A longitudinal MR imaging study. *American Journal of Neuroradiology*, 31(6), pp.1036–1041. Available at: http://www.ajnr.org/content/ajnr/31/6/1036.full.pdf [Accessed January 3, 2018].

Hobbs, N.Z. et al., 2015. Short-interval observational data to inform clinical trial design in Huntington's disease. *Journal of neurology, neurosurgery, & psychiatry*, 86(12), pp.1291–8. Available at: http://jnnp.bmj.com/cgi/doi/10.1136/jnnp-2014-309768 [Accessed January 19, 2016].

Hobbs, N.Z., Henley, S.M.D., et al., 2010. The progression of regional atrophy in premanifest and early Huntington's disease: a longitudinal voxel-based morphometry study. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(7), pp.756–763. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19955112 [Accessed May 8, 2017].

Huntington Study Group, 1996. Unified Huntington's Disease Rating Scale: reliability and consistency. *Mov Disord.*, 11(2), pp.136–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8684382.

Hupé, J.-M., 2015. Statistical inferences under the Null hypothesis: common mistakes and pitfalls in neuroimaging studies. *Frontiers in neuroscience*, 9, p.18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25745383 [Accessed September 26, 2017].

Irimia, A. et al., 2012. Neuroimaging of structural pathology and connectomics in traumatic brain injury: Toward personalized outcome prediction. *NeuroImage. Clinical*, 1(1), pp.1–17. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3757727&tool=pmcentrez&rendertype=abstract [Accessed November 20, 2015].

Iscan, Z. et al., 2015. Test-retest reliability of freesurfer measurements within and between sites: Effects of visual approval process. *Human brain mapping*, 36(9), pp.3472–85.

Available at: http://www.ncbi.nlm.nih.gov/pubmed/26033168 [Accessed December 11, 2015].

Jahanshahi, M. et al., 2015. A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. *Nature Reviews Neuroscience*, 16(12), pp.719–732.

Jenkinson, M. et al., 2012. FSL. *NeuroImage*, 62, pp.782–90.

Jenkinson, M. et al., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17, pp.825–841.

Jenkinson, M. & Smith, S., 2001. A global optimisation method for robust affine registration of brain images. *Med.Image Anal.*, 5, pp.143–156.

Johnson, E.B. et al., 2018. Neurofilament light protein in blood predicts regional atrophy in Huntington disease. *Neurology*, 90(8), pp.e717–e723. Available at: http://www.neurology.org/lookup/doi/10.1212/WNL.0000000000005005 [Accessed February 22, 2018].

Johnson, E.B. et al., 2017. Recommendations for the Use of Automated Gray Matter Segmentation Tools: Evidence from Huntington's Disease. *Frontiers in neurology*, 8, p.519. Available at: http://journal.frontiersin.org/article/10.3389/fneur.2017.00519/full [Accessed November 1, 2017].

Johnson, E.B. et al., 2015. The impact of occipital lobe cortical thickness on cognitive task performance: An investigation in Huntington's Disease. *Neuropsychologia*, 79, pp.138–146.

Jones, C.K. et al., 2006. Amide proton transfer imaging of human brain tumors at 3T. *Magnetic resonance in medicine*, 56(3), pp.585–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16892186.

Jovicich, J. et al., 2013. Brain morphometry reproducibility in multi-center 3T MRI studies: A comparison of cross-sectional and longitudinal segmentations. *NeuroImage*, 83, pp.472–484.

Juan, F.D. et al., 2016. Multimodal imaging biomarkers in premanifest and early Huntington's disease: 30-month IMAGE-HD data. *British Journal of Psychiatry*, 208(6), pp.571–578.

Kassubek, J. et al., 2004. Topography of cerebral atrophy in early Huntington's disease: a voxel based morphometric MRI study. *Journal of neurology, neurosurgery, and psychiatry*, 75(2), pp.213–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14742591 [Accessed November 29, 2017].

Katuwal, G.J. et al., 2016. Inter-Method Discrepancies in Brain Volume Estimation May Drive Inconsistent Findings in Autism. *Frontiers in neuroscience*, 10, p.439. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27746713 [Accessed January 25, 2017].

Kazemi, K. & Noorizadeh, N., 2014. Quantitative Comparison of SPM, FSL, and Brainsuite for Brain MR Image Segmentation. *Journal of biomedical physics & engineering*, 4(1), pp.13–26. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25505764 [Accessed August 31, 2016].

Keller, S.S. et al., 2004. Comparison of standard and optimized voxel-based morphometry for analysis of brain changes associated with temporal lobe epilepsy. *NeuroImage*, 23(3), pp.860–868.

Klauschen, F. et al., 2009. Evaluation of automated brain MR image segmentation and volumetry methods. *Human brain mapping*, 30(4), pp.1310–27. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/18537111 [Accessed January 26, 2016].

Klöppel, S. et al., 2015. Compensation in Preclinical Huntington's Disease: Evidence From the Track-On HD Study. *EBioMedicine*, 2(10), pp.1420–9. Available at: http://www.sciencedirect.com/science/article/pii/S2352396415300967 [Accessed January 18, 2016].

Kong, L. et al., 2012. Reduced Gray to White Matter Tissue Intensity Contrast in Schizophrenia G. M. McAlonan, ed. *PLoS ONE*, 7(5), p.e37016. Available at: http://dx.plos.org/10.1371/journal.pone.0037016 [Accessed October 11, 2016].

Koo, T.K. & Li, M.Y., 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, 15(2), pp.155–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27330520 [Accessed December 5, 2017].

De La Monte, S.M., Vonsattel, J.P. & Richardson, E.P., 1988. Morphometric demonstration of atrophic changes in the cerebral cortex, white matter, and neostriatum in Huntington's disease. *Journal of neuropathology and experimental neurology*, 47(5), pp.516–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2971785 [Accessed November 21, 2017].

Langbehn, D.R. et al., 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics*, 65(4), pp.267–277. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15025718 [Accessed January 8, 2018].

Langbehn, D.R. et al., 2010. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. *Am.J Med.Genet.B Neuropsychiatr.Genet.*, 153B(2), pp.397–408. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19548255 [Accessed May 8, 2017].

Ledig, C. et al., 2015. Robust whole-brain segmentation: application to traumatic brain injury. *Medical image analysis*, 21(1), pp.40–58. Available at: http://www.sciencedirect.com/science/article/pii/S136184151400187X [Accessed December 22, 2015].

Lee, J. et al., 2006. A novel quantitative validation of the cortical surface reconstruction algorithm using MRI phantom: issues on local geometric accuracy and cortical thickness . *Med.Image Comput.Comput.Assist.Interv.* , 9 (Pt 1), pp.183–190.

Liem, F. et al., 2015. Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *NeuroImage*, 108, pp.95–109. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25534113 [Accessed April 5, 2017].

Lindquist, M.A. & Mejia, A., 2015. Zen and the art of multiple comparisons. *Psychosomatic medicine*, 77(2), pp.114–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25647751 [Accessed September 26, 2017].

Littmann, A. et al., 2006. Acquisition-Related Morphological Variability in Structural MRI 1. *Academic Radiology*, 13(9), pp.1055–1061. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16935717 [Accessed January 12, 2018].

MacDonald, M.E. et al., 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*, 72(6), pp.971–83. Available at: http://www.sciencedirect.com/science/article/pii/009286749390585E?via%3Dihub [Accessed October 17, 2017].

MacLeod, R. et al., 2013. Recommendations for the predictive genetic test in Huntington's disease. *Clinical Genetics*, 83(3), pp.221–231. Available at: http://doi.wiley.com/10.1111/j.1399-0004.2012.01900.x [Accessed October 22, 2017].

Magnotta, V.A. et al., 2002. Structural MR image processing using the BRAINS2 toolbox. *Comput.Med.Imaging Graph.*, 26, pp.251–264.

Majid, D.S.A. et al., 2011. Basal ganglia atrophy in prodromal Huntington's disease is detectable over one year using automated segmentation. *Movement Disorders*, 26(14), pp.2544–2551.

Malone, I.B. et al., 2015. Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance. *NeuroImage*, 104, pp.366–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25255942 [Accessed September 20, 2017].

Mann, D.M., Oliver, R. & Snowden, J.S., 1993. The topographic distribution of brain atrophy in Huntington's disease and progressive supranuclear palsy. *Acta neuropathologica*, 85(5), pp.553–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8493863 [Accessed November 21, 2017].

Matsumae, M. et al., 1996. Age-related changes in intracranial compartment volumes in normal adults assessed by magnetic resonance imaging. *Journal of Neurosurgery*, 84(6), pp.982–991. Available at: http://thejns.org/doi/10.3171/jns.1996.84.6.0982 [Accessed January 10, 2018].

Mazziotta, J. et al., 2001. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos.Trans.R.Soc.Lond B Biol.Sci.*, 356, pp.1293–1322.

McAuliffe, M.J. et al., 2001. Medical Image Processing, Analysis & Visualization in Clinical Research. , p.381. Available at: http://dl.acm.org/citation.cfm?id=872745.873034 [Accessed January 18, 2016].

McCarthy, C.S. et al., 2015. A comparison of FreeSurfer-generated data with and without manual intervention. , 9(OCT).

McColgan, P. et al., 2015. Selective vulnerability of Rich Club brain regions is an organizational principle of structural connectivity loss in Huntington's disease. *Brain*, 138(Pt 11), pp.3327–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26384928 [Accessed May 10, 2017].

McColgan, P. et al., 2017. Topological length of white matter connections predicts their rate of atrophy in premanifest Huntington's disease. *JCI Insight*, 2(8). Available at: http://www.ncbi.nlm.nih.gov/pubmed/28422761 [Accessed August 3, 2017].

McCusker, E.A. & Loy, C.T., 2017. Huntington Disease: The Complexities of Making and Disclosing a Clinical Diagnosis After Premanifest Genetic Testing. *Tremor and other hyperkinetic movements (New York, N.Y.)*, 7, p.467. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28975045 [Accessed January 3, 2018].

McIntosh, A.R. & Misic, B., 2013. Multivariate statistical analyses for neuroimaging data. *Annual Review of Psychology*, 64(1), pp.499–525. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22804773 [Accessed November 2, 2017].

Minkova, L. et al., 2018. Cross-sectional and longitudinal voxel-based grey matter asymmetries in Huntington's disease. *NeuroImage: Clinical*, 17, pp.312–324.

Minkova, L. et al., 2017. Gray matter asymmetries in aging and neurodegeneration: A review

and meta-analysis. *Human Brain Mapping*, 38(12), pp.5890–5904.

Morrison, P., Harding-Lester, S. & Bradley, A., 2011. Uptake of Huntington disease predictive testing in a complete population. *Clinical Genetics*, 80(3), pp.281–286. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20880124 [Accessed October 17, 2017].

Moss, D.J.H. et al., 2017. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *The Lancet Neurology*, 16(9), pp.701–711. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28642124 [Accessed January 3, 2018].

Nana, A.L. et al., 2014. Widespread heterogeneous neuronal loss across the cerebral cortex in Huntington's disease. *Journal of Huntington's disease*, 3(1), pp.45–64. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25062764 [Accessed November 14, 2017].

Nopoulos, P.C. et al., 2010. Cerebral cortex structure in prodromal Huntington disease. *Neurobiol.Dis.*, 40, pp.544–554.

Nopoulos, P.C. et al., 2011. Smaller intracranial volume in prodromal Huntington's disease: Evidence for abnormal neurodevelopment. *Brain*, 134, pp.137–142. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3025719&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2014].

Ogawa, S. et al., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24), pp.9868–72. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=55275&tool=pmcentrez&rendertype=abstract.

Panizzon, M.S. et al., 2009. Distinct genetic influences on cortical surface area and cortical thickness. *Cerebral cortex (New York, N.Y. : 1991)*, 19(11), pp.2728–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19299253 [Accessed April 3, 2017].

Papoutsi, M. et al., 2014. The cognitive burden in Huntington's disease: Pathology, phenotype, and mechanisms of compensation. *Movement Disorders*, 29(5), pp.673–683. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24757115 [Accessed October 22, 2017].

Papp, K. V, Kaplan, R.F. & Snyder, P.J., 2011. Biological markers of cognition in prodromal Huntington's disease: a review. *Brain Cogn*, 77, pp.280–291.

Paulsen, J.S., 2009. Biomarkers to predict and track diseases. *The Lancet. Neurology*, 8(9), pp.776–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19646925 [Accessed November 26, 2017].

Paulsen, J.S. et al., 2006. Brain structure in preclinical Huntington's disease. *Biological psychiatry*, 59(1), pp.57–63. Available at: http://www.sciencedirect.com/science/article/pii/S0006322305007092 [Accessed July 8, 2015].

Paulsen, J.S., Long, J.D., Johnson, H.J., et al., 2014. Clinical and Biomarker Changes in Premanifest Huntington Disease Show Trial Feasibility: A Decade of the PREDICT-HD Study. *Front Aging Neurosci.*, 6, p.78.

Paulsen, J.S. et al., 2008. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *J Neurol Neurosurg Psychiatry*, 79(8), pp.874–880.

Paulsen, J.S., Long, J.D., Ross, C.A., et al., 2014. Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet*

Neurology, 13(12), pp.1193–1201. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1474442214702388 [Accessed November 17, 2014].

Paulsen, J.S. et al., 2010. Striatal and white matter predictors of estimated diagnosis for Huntington disease. *Brain research bulletin*, 82(3-4), pp.201–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2892238&tool=pmcentrez&rendertype=abstract [Accessed November 7, 2015].

Penney, J.B. et al., 1997. CAG repeat number governs the development rate of pathology in Huntington's disease. *Annals of neurology*, 41(5), pp.689–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9153534 [Accessed October 11, 2015].

Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, 59(1), pp.319–330.

Perlaki, G. et al., 2017. Comparison of accuracy between FSL's FIRST and Freesurfer for caudate nucleus and putamen segmentation. *Scientific Reports*, 7(1), p.2418. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28546533 [Accessed July 17, 2017].

Raj, A. et al., 2015. Network Diffusion Model of Progression Predicts Longitudinal Patterns of Atrophy and Metabolism in Alzheimer's Disease. *Cell Reports*, 10(3), pp.359–369.

Raj, A., Kuceyeski, A. & Weiner, M., 2012. A Network Diffusion Model of Disease Progression in Dementia. *Neuron*, 73(6), pp.1204–1215.

Read, J. et al., 2013. Quality of life in huntington's disease: A comparative study investigating the impact for those with pre-manifest and early manifest disease, and their partners. *Journal of Huntington's Disease*, 2(2), pp.159–175.

Ready, R.E. et al., 2008. Patient and caregiver quality of life in Huntington's disease. *Movement Disorders*, 23(5), pp.721–726. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-44449146417&partnerID=40&md5=1032d3a6d4d086a6e14e178d8dcd9d67.

Rees, E., 2014. Development and evaluation of biomarkers in Huntington's Disease: furthering our understanding of the disease and preparing for clinical trials. *Doctoral thesis, UCL (University College London).* . Available at: http://discovery.ucl.ac.uk/1455986/ [Accessed February 16, 2018].

Righart, R. et al., 2017. Volume versus surface-based cortical thickness measurements: A comparative study with healthy controls and multiple sclerosis patients F. Paul, ed. *PLOS ONE*, 12(7), p.e0179590. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28683072 [Accessed July 17, 2017].

Rodrigues, F.B. & Wild, E.J., 2017. Clinical Trials Corner: September 2017. *Journal of Huntington's Disease*, 6(3), pp.255–263. Available at: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JHD-170262 [Accessed November 25, 2017].

Rosas, H.D. et al., 2011. A tale of two factors: what determines the rate of progression in Huntington's disease? A longitudinal MRI study. *Mov Disord.*, 26, pp.1691–1697.

Rosas, H.D. et al., 2008. Cerebral cortex and the clinical expression of Huntington's disease: complexity and heterogeneity. *Brain*, 131, pp.1057–1068.

Rosas, H.D. et al., 2002. Regional and progressive thinning of the cortical ribbon in Huntington's disease. *Neurology*, 58, pp.695–701.

Rosenblatt, A. et al., 2012. Age, CAG repeat length, and clinical progression in Huntington's disease. *Movement Disorders*, 27(2), pp.272–276. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22173986 [Accessed January 12, 2018].

Rosenblatt, A. et al., 2006. The association of CAG repeat length with clinical progression in Huntington disease. *Neurology*, 66(7), pp.1016–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8208412 [Accessed January 9, 2018].

Ross, C. a et al., 2014. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nature reviews. Neurology*, 10(4), pp.204–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24614516.

Rüb, U. et al., 2015. Huntington's disease (HD): neurodegeneration of Brodmann's Primary Visual Area 17 (BA17). *Brain pathology*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25495445 [Accessed February 26, 2015].

Ruocco, H.H. et al., 2008. Longitudinal analysis of regional grey matter loss in Huntington disease: Effects of the length of the expanded CAG repeat. *Journal of Neurology, Neurosurgery and Psychiatry*, 79(2), pp.130–135. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14742591 [Accessed January 3, 2018].

Scahill, R.I. et al., 2003. A Longitudinal Study of Brain Volume Changes in Normal Aging Using Serial Registered Magnetic Resonance Imaging. *Archives of Neurology*, 60(7), p.989. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12873856 [Accessed September 20, 2017].

Scahill, R.I. et al., 2013. Clinical impairment in premanifest and early Huntington's disease is associated with regionally specific atrophy. *Human Brain Mapping*, 34(3), pp.519–529.

Scahill, R.I., Wild, E.J. & Tabrizi, S.J., 2012. Biomarkers for Huntington's disease: an update. *Expert Opinion on Medical Diagnostics*, 6(5), pp.371–375. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23480802 [Accessed November 27, 2017].

Schnack, H.G. et al., 2010. Mapping reliability in multicenter MRI: Voxel-based morphometry and cortical thickness. *Human Brain Mapping*, 31(12), pp.1967–1982. Available at: http://doi.wiley.com/10.1002/hbm.20991 [Accessed April 5, 2017].

Schwarz, C.G. et al., 2016. A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. *NeuroImage. Clinical*, 11, pp.802–812. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28050342 [Accessed April 5, 2017].

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, pp.87–97. Available at: file://m/ION/TrackHD/Imaging/Eli/database/A_nonparametric_method_for_automatic_correction.pdf.

Smith, A., 1991. *Symbol Digit Modalities Test*, Los Angeles: Western Psychological Services.

Smith, S.M., 2002. Fast robust automated brain extraction. *Human brain mapping*, 17(3), pp.143–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12391568 [Accessed September 29, 2016].

Sormani, M.P. et al., 2004. Measurement error of two different techniques for brain atrophy assessment in multiple sclerosis. *Neurology*, 62(8), pp.1432–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15111692 [Accessed November 30, 2017].

Stephan, K.E. et al., 2007. Dynamic causal models of neural system dynamics:current state and

future extensions. *Journal of biosciences*, 32(1), pp.129–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17426386 [Accessed December 14, 2017].

Stephan, K.E., Penny, W.D., Moran, R.J., den Ouden, H.E.M., Daunizeau, J., Friston, K.J., 2010. Ten simple rules for dynamic causal modeling. Neuroimage 49, 3099–3109. doi:10.1016/j.neuroimage.2009.11.015

Stoffers, D. et al., 2010. Contrasting gray and white matter changes in preclinical Huntington disease: An MRI study. *Neurology*, 74(15), pp.1208–1216.

Storsve, A.B. et al., 2014. Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *J Neurosci.*, 34, pp.8488–8498.

Tabrizi, S.J. et al., 2011. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol*, 10, pp.31–42.

Tabrizi, S.J. et al., 2009. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *The Lancet Neurology*, 8(9), pp.791–801. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19646924 [Accessed October 28, 2014].

Tabrizi, S.J. et al., 2012. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *The Lancet Neurology*, 11(1), pp.42–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22137354 [Accessed September 19, 2014].

Tabrizi, S.J. et al., 2013. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol*, 12, pp.637–649.

Thambisetty, M. et al., 2010. Longitudinal changes in cortical thickness associated with normal aging. *NeuroImage*, 52(4), pp.1215–1223. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20441796 [Accessed April 3, 2017].

The U.S.-Venezuela Collaborative, N.S. et al., 2004. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proceedings of the National Academy of Sciences*, 101(10), pp.3498–3503. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14993615 [Accessed October 17, 2017].

Thieben, M.J. et al., 2002. The distribution of structural neuropathology in pre-clinical Huntington's disease. *Brain*, 125(8), pp.1815–1828. Available at: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awf179 [Accessed November 29, 2017].

Thu, D.C. V. et al., 2010. Cell loss in the motor and cingulate cortex correlates with symptomatology in Huntington's disease. *Brain*, 133(4), pp.1094–1110. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20375136 [Accessed November 21, 2017].

Tibben, A., 2007. Predictive testing for Huntington's disease. *Brain Research Bulletin*, 72(2-3), pp.165–171. Available at: https://www.sciencedirect.com/science/article/pii/S0361923006003297 [Accessed October 22, 2017].

Tohka, J., 2014. Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review. *World journal of radiology*, 6(11), pp.855–64. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4241492&tool=pmcentrez&rendertype=abstract [Accessed January 23, 2016].

Tustison, N.J. et al., 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage*, 99, pp.166–179.

Tustison, N.J. et al., 2010. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6), pp.1310–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3071855&tool=pmcentrez&rendertype=abstract [Accessed October 5, 2015].

UNESCO, 1997. *International standard classification of education ISCED 1997*, Available at: www.uis.unesco.org.

Vonsattel, J.P. et al., 1985. Neuropathological classification of Huntington's disease. *Journal of neuropathology and experimental neurology*, 44(6), pp.559–77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2932539 [Accessed November 21, 2017].

Waters, A. et al., 2017. Using Outliers in Freesurfer Segmentation Statistics to Identify Cortical Reconstruction Errors in Structural Scans. *bioRxiv*. Available at: https://www.biorxiv.org/content/early/2017/08/21/176818 [Accessed September 26, 2017].

Whitwell, J.L. et al., 2001. Normalization of Cerebral Volumes by Use of Intracranial Volume: Implications for Longitudinal Quantitative MR Imaging. *AJNR Am J Neuroradiol*, 22, pp.1483–1489. Available at: http://www.ajnr.org/content/ajnr/22/8/1483.full.pdf [Accessed January 10, 2018].

Wild, E.J. & Tabrizi, S.J., 2017. Therapies targeting DNA and RNA in Huntington's disease. *The Lancet Neurology*, 16(10), pp.837–847. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1474442217302806 [Accessed September 15, 2017].

Winkler, A. et al., 2009. Heritability of Volume, Surface Area and Thickness for Anatomically Defined Cortical Brain Regions Estimated in a Large Extended Pedigree. *NeuroImage*, 47. Available at: http://www.sciencedirect.com/science/article/pii/S1053811909717136 [Accessed April 3, 2017].

Winkler, A.M. et al., 2010. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*, 53, pp.1135–1146.

Wolf, R.C. et al., 2009. Cortical dysfunction in patients with Huntington's disease during working memory performance. *Human Brain Mapping*, 30(1), pp.327–339. Available at: http://doi.wiley.com/10.1002/hbm.20502 [Accessed November 30, 2017].

Wolf, R.C. et al., 2014. Visual system integrity and cognition in early Huntington's disease. *Eur.J Neurosci.*, 40(2), pp.2417–2426.

Wonderlick, J.S. et al., 2009. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. *NeuroImage*, 44, pp.1324–1333.

Wu, D. et al., 2017. Mapping the order and pattern of brain structural MRI changes using change-point analysis in premanifest Huntington's disease. *Human Brain Mapping*, 38(10), pp.5035–5050. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28657159 [Accessed November 2, 2017].

Xu, L. et al., 2009. Source-based morphometry: the use of independent component analysis to

identify gray matter differences with application to schizophrenia. *Human brain mapping*, 30(3), pp.711–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18266214 [Accessed November 30, 2017].

Zhang, H. et al., 2012. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage*, 61(4), pp.1000–1016.

Zhang, Y. et al., 2011. Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 156B(7), pp.751–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21858921 [Accessed January 15, 2018].

Zhang, Y., Brady, M. & Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans.Med.Imaging*, 20, pp.45–57.

Zhou, J. et al., 2012. Predicting Regional Neurodegeneration from the Healthy Brain Functional Connectome. *Neuron*, 73(6), pp.1216–1227.

Ziegler, G. et al., 2017. Multivariate dynamical modelling of structural change during development. *NeuroImage*, 147, pp.746–762. Available at: http://linkinghub.elsevier.com/retrieve/pii/S105381191630739X [Accessed March 27, 2017].

Zou, K.H. et al., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic radiology*, 11(2), pp.178–89. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14974593 [Accessed September 26, 2017].

9. APPENDICES

Appendix 1

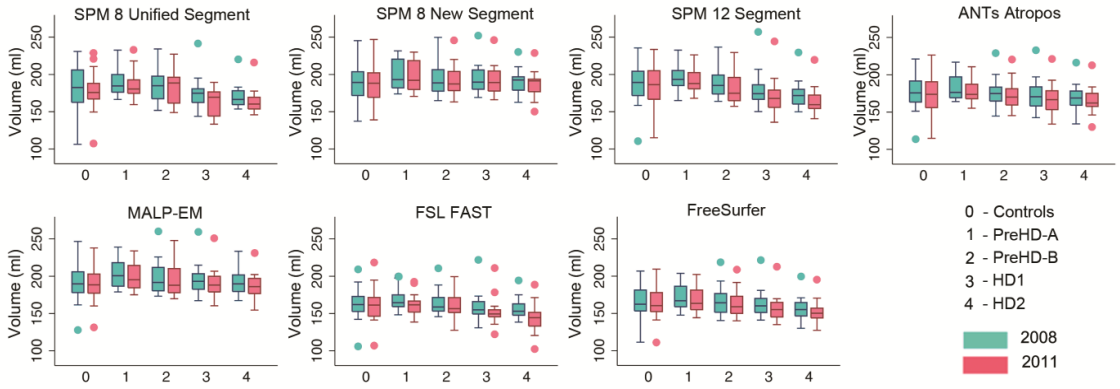


Figure 1: Box plots showing frontal GM Volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

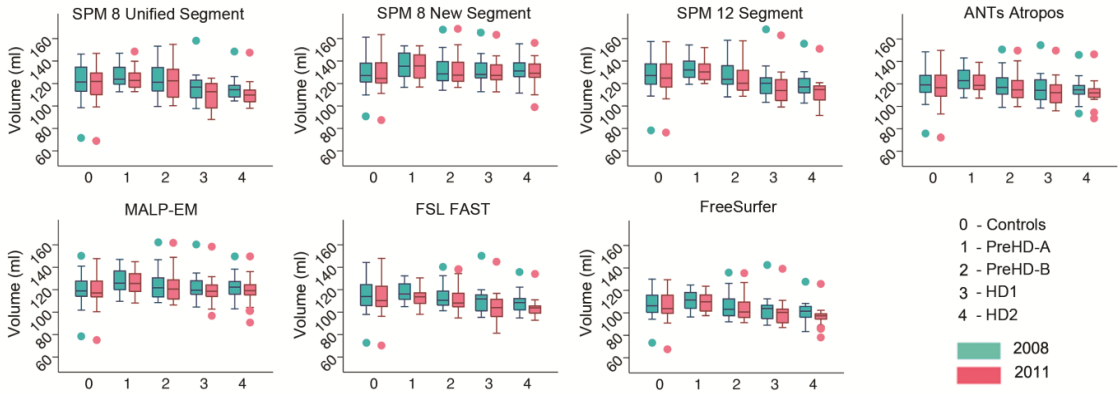


Figure 2: Box plots showing temporal GM Volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

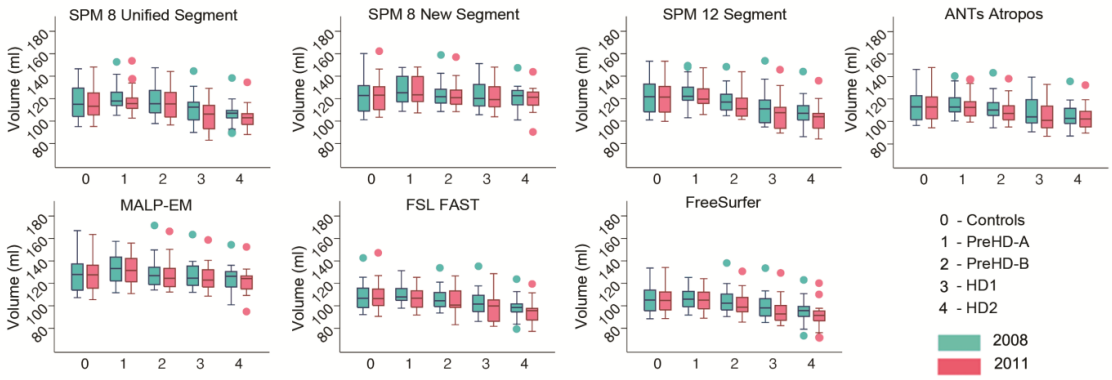


Figure 3: Box plots showing parietal GM Volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

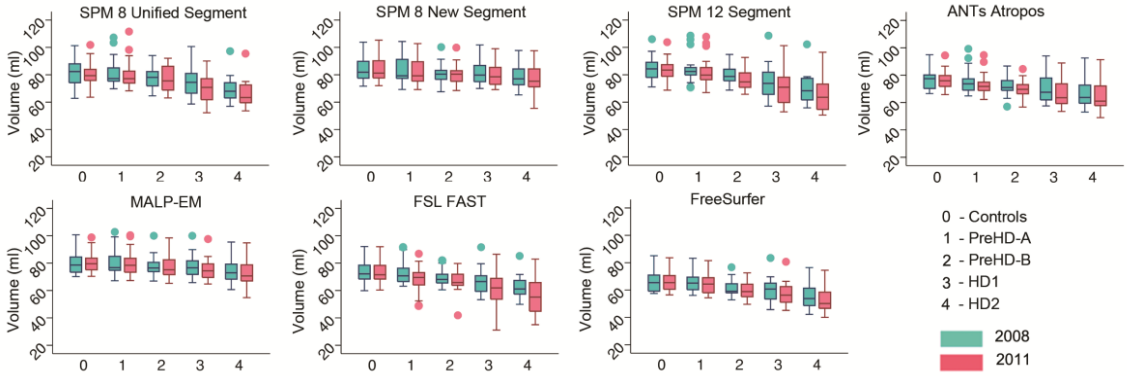


Figure 4: Box plots showing occipital GM Volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

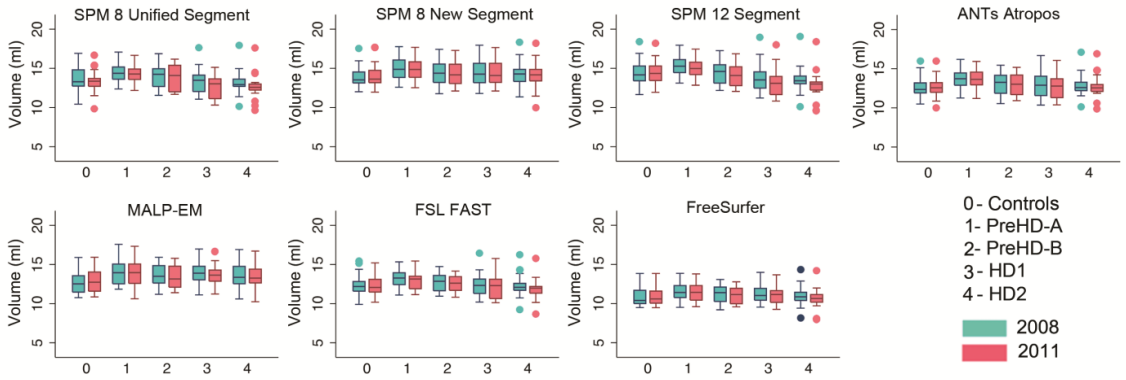


Figure 5: Box plots showing insula GM Volumes for all groups and all tools for 2008 and 2011 time points. Boxes show the first quartile, median and third quartile, with whiskers representing the smallest and largest value not classified as an outlier. Dots represent outliers.

Table 1: Intraclass correlations and confidence intervals for all HD groups in frontal, temporal, parietal, occipital lobes and the insula in different tools.

	PreHD-A (N=20)	PreHD-B (N=20)	HD1 (N=20)	HD2 (N=20)
		Frontal GM Intraclass Correlations Confidence intervals		
SPM 8 Unified	0.991 0.977-0.997	0.979 0.945-0.992	0.986 0.967-0.995	0.990 0.976-0.996
SPM 8 New Segment	0.997 0.994-0.999	0.999 0.994-1.000	0.999 0.997-1.000	0.998 0.994-0.999
SPM 12	0.987 0.969-0.995	0.992 0.981-0.997	0.998 0.994-0.999	0.994 0.983-0.998
Atropos	0.957 0.868-0.984	0.995 0.987-0.998	0.990 0.970-0.996	0.989 0.952-0.997
MALP-EM	0.992 0.981-0.997	0.998 0.995-0.999	0.998 0.996-0.999	0.996 0.990-0.998
FAST	0.983 0.959-0.993	0.993 0.984-0.997	0.995 0.981-0.998	0.991 0.977-0.996
FreeSurfer	0.973 0.927-0.990	0.977 0.942-0.991	0.984 0.949-0.994	0.988 0.971-0.995
		Temporal GM Intraclass Correlations Confidence intervals		
SPM 8 Unified	0.992 0.979-0.997	0.979 0.947-0.992	0.990 0.976-0.996	0.993 0.982-0.997
SPM 8 New Segment	0.999 0.996-0.999	0.999 0.998-1.000	0.998 0.996-0.999	0.998 0.996-0.999
SPM 12	0.987 0.965-0.995	0.992 0.979-0.997	0.997 0.992-0.999	0.995 0.988-0.998
Atropos	0.959 0.847-0.986	0.994 0.983-0.998	0.995 0.985-0.998	0.991 0.959-0.997
MALP-EM	0.991 0.962-0.997	0.998 0.995-0.999	0.997 0.993-0.999	0.994 0.984-0.997
FAST	0.970 0.903-0.989	0.993 0.981-0.997	0.996 0.991-0.999	0.994 0.986-0.998
FreeSurfer	0.959 0.894-0.984	0.979 0.945-0.992	0.995 0.987-0.998	0.987 0.968-0.995
		Parietal GM Intraclass Correlations Confidence intervals		
SPM 8 Unified	0.994 0.984-0.998	0.984 0.959-0.994	0.988 0.970-0.995	0.990 0.976-0.996
SPM 8 New Segment	0.999 0.997-1.000	0.994 0.986-0.998	0.998 0.995-0.999	0.998 0.995-0.999
SPM 12	0.988 0.970-0.995	0.992 0.981-0.997	0.998 0.994-0.999	0.994 0.985-0.998
Atropos	0.987 0.952-0.996	0.992 0.978-0.997	0.989 0.968-0.996	0.986 0.948-0.995
MALP-EM	0.997 0.993-0.999	0.994 0.984-0.998	0.999 0.997-1.000	0.996 0.989-0.998
FAST	0.989 0.973-0.996	0.992 0.980-0.997	0.990 0.971-0.996	0.992 0.981-0.997
FreeSurfer	0.991 0.977-0.996	0.978 0.936-0.992	0.981 0.945-0.993	0.987 0.967-0.995
		Occipital GM Intraclass Correlations Confidence intervals		

SPM 8 Unified	0.993 0.982-0.997	0.968 0.920-0.988	0.983 0.958-0.993	0.991 0.978-0.996
SPM 8 New Segment	0.999 0.997-1.000	0.993 0.983-0.997	0.998 0.994-0.999	0.996 0.991-0.999
SPM 12	0.990 0.976-0.996	0.973 0.934-0.989	0.998 0.994-0.999	0.993 0.979-0.997
Atropos	0.984 0.947-0.994	0.978 0.913-0.993	0.991 0.977-0.996	0.995 0.967-0.998
MALP-EM	0.994 0.985-0.998	0.991 0.974-0.996	0.994 0.985-0.998	0.996 0.991-0.999
FAST	0.988 0.971-0.995	0.990 0.958-0.997	0.995 0.986-0.998	0.997 0.992-0.999
FreeSurfer	0.987 0.967-0.995	0.968 0.908-0.988	0.989 0.973-0.996	0.995 0.988-0.981

		Insula **Intraclass Correlation** **Confidence Intervals**		
SPM 8 Unified	0.990 0.974-0.996	0.984 0.960-0.994	0.994 0.984-0.997	0.994 0.984-0.998
SPM 8 New Segment	0.995 0.988-0.998	0.996 0.991-0.999	0.998 0.996-0.999	0.998 0.995-0.999
SPM 12	0.988 0.969-0.995	0.987 0.968-0.995	0.997 0.993-0.999	0.996 0.989-0.998
Atropos	0.988 0.956-0.996	0.993 0.984-0.997	0.995 0.987-0.998	0.993 0.981-0.997
MALP-EM	0.996 0.990-0.999	0.998 0.994-0.999	0.997 0.992-0.999	0.996 0.990-0.998
FAST	0.992 0.973-0.997	0.995 0.988-0.998	0.997 0.993-0.999	0.997 0.992-0.999
FreeSurfer	0.984 0.960-0.994	0.970 0.925-0.988	0.964 0.906-0.986	0.963 0.910-0.985)

Table 2: Repeatability values for back-to-back segmentations of frontal, temporal, parietal, occipital and insula GM for all HD participants included in the current study, showing means, standard deviations, and ranges.

	PreHD-A (N=20)	PreHD-B (N=20)	HD1 (N=20)	HD2 (N=20)
	Frontal			
	Mean repeatability (Standard Deviation)			
	Range			
SPM 8 Unified Segment	1.06 (0.88) 0.14-3.59	1.68 (2.09) 0.05-8.02	1.11 (1.60) 0.11-7.42	0.99 (0.78) 0.03-2.76
SPM 8 New Segment	0.55 (0.45) 0.01-1.44	0.45 (0.32) 0.00-1.09	0.40 (0.25) 0.05-0.94	0.44 (0.33) 0.04-1.30
SPM 12	1.12 (1.07) 0.08-4.27	0.89 (0.78) 0.04-2.35	0.71 (0.60) 0.00-1.93	0.91 (0.69) 0.20-3.05
Atropos	1.72 (2.37) 0.05-9.03	0.84 (0.64) 0.08-2.48	1.27 (1.18) 0.04-4.02	1.22 (1.13) 0.02-4.16
MALP-EM	0.76 (0.93) 0.00-3.92	0.56 (0.39) 0.02-1.33	0.43 (0.37) 0.01-1.58	0.57 (0.60) 0.01-2.01
FAST	1.22 (0.80) 0.16-2.95	0.93 (0.67) 0.06-2.43	0.97 (0.77) 0.19-2.78	0.93 (0.87) 0.03-2.69
FreeSurfer	1.43 (1.71) 0.00-5.29	3.21 (5.54) 0.20-26.01	1.59 (1.35) 0.02-5.51	1.28 (1.02) 0.10-3.43
	Temporal			
	Mean repeatability (Standard Deviation)			
	Range			
SPM 8 Unified Segment	0.80 (0.76) 0.02-2.66	1.49 (1.92) 0.17-7.06	1.11 (1.05) 0.19-4.97	0.85 (0.72) 0.03-2.57
SPM 8 New Segment	0.37 (0.26) 0.03-0.96	0.34 (0.29) 0.03-0.86	0.38 (0.34) 0.04-1.33	0.36 (0.33) 0.03-1.08
SPM 12	0.98 (0.74) 0.10-2.81	0.81 (0.84) 0.06-3.22	0.74 (0.58) 0.06-2.38	0.81 (0.55) 0.10-2.01
Atropos	1.43 (1.68) 0.17-5.97	0.73 (0.79) 0.03-2.85	0.75 (0.78) 0.03-3.33	1.07 (1.03) 0.15-3.61
MALP-EM	0.77 (0.77) 0.01-2.97	0.48 (0.40) 0.00-1.67	0.52 (0.60) 0.00-2.64	0.87 (0.78) 0.07-2.36
FAST	1.02 (1.20) 0.03-5.12	0.75 (0.78) 0.03-2.95	0.69 (0.58) 0.02-2.13	0.69 (0.77) 0.02-3.26
FreeSurfer	3.15 (2.30) 0.18-7.87	3.04 (3.51) 0.18-15.82	3.43 (2.33) 0.26-8.30	4.51 (2.81) 0.09-8.49
	Parietal			
	Mean repeatability (Standard Deviation)			
	Range			
SPM 8 Unified Segment	0.89 (0.83) 0.22-3.39	1.57 (1.64) 0.02-6.21	1.15 (1.37) 0.01-6.38	1.04 (0.81) 0.04-2.66
SPM 8 New Segment	0.38 (0.33) 0.02-1.49	0.54 (0.82) 0.00-3.79	0.46 (0.37) 0.00-1.44	0.39 (0.39) 0.00-1.33
SPM 12	1.10 (1.02) 0.01-3.89	0.84 (0.78) 0.04-2.83	0.71 (0.65) 0.05-2.42	0.90 (0.73) 0.07-3.40
Atropos	1.08 (1.12) 0.10-4.16	0.98 (0.91) 0.04-3.15	1.22 (1.26) 0.07-4.43	1.35 (1.45) 0.09-5.60
MALP-EM	0.50 (0.53) 0.05-2.32	0.72 (0.89) 0.03-3.79	0.37 (0.30) 0.02-1.29	0.72 (0.43) 0.14-1.73
FAST	0.95 (0.75) 0.09-2.50	0.94 (0.71) 0.03-2.43	1.31 (1.04) 0.12-3.14	0.82 (0.90) 0.05-3.66
FreeSurfer	0.90 (0.83) 0.00-2.63	1.99 (1.72) 0.40-7.01	1.66 (1.70) 0.02-6.43	1.32 (1.43) 0.01-4.14
	Occipital			
	Mean repeatability (Standard Deviation)			

		Range		
SPM 8 Unified Segment	1.00 (1.24) 0.25-5.84	2.03 (2.01) 0.04-7.31	1.55 (1.84) 0.05-8.44	1.38 (1.20) 0.10-3.91
SPM 8 New Segment	0.52 (0.31) 0.02-1.07	0.82 (0.61) 0.01-2.36	0.58 (0.44) 0.01-1.45	0.67 (0.60) 0.02-2.10
SPM 12	1.27 (1.02) 0.03-4.33	1.34 (1.32) 0.00-4.65	0.90 (0.71) 0.02-2.18	1.49 (1.20) 0.32-4.53
Atropos	1.56 (1.68) 0.15-6.95	1.44 (1.40) 0.02-6.44	1.19 (1.27) 0.00-4.47	1.25 (1.15) 0.07-4.49
MALP-EM	1.05 (0.67) 0.05-2.82	0.98 (0.91) 0.02-3.99	0.84 (0.73) 0.03-2.91	0.80 (0.72) 0.06-2.42
FAST	1.39 (1.00) 0.14-4.15	0.94 (0.80) 0.10-2.57	1.17 (0.73) 0.13-2.63	0.86 (0.79) 0.11-3.72
FreeSurfer	1.47 (1.24) 0.17-4.96	2.25 (2.02) 0.13-7.06	1.59 (1.35) 0.09-5.90	1.14 (1.13) 0.09-4.54
		Insula **Mean repeatability (Standard Deviation)** **Range**		
SPM 8 Unified Segment	0.82 (1.00) 0.01-3.21	1.43 (1.36) 0.06-4.45	1.02 (0.99) 0.01-4.53	1.03 (0.93) 0.08-3.39
SPM 8 New Segment	0.81 (0.63) 0.02-2.08	0.80 (0.44) 0.09-1.56	0.60 (0.38) 0.02-1.33	0.53 (0.43) 0.08-1.37
SPM 12	0.86 (0.95) 0.11-3.88	1.08 (1.04) 0.09-4.00	0.75 (0.67) 0.04-2.65	0.97 (0.84) 0.02-3.36
Atropos	1.26 (0.73) 0.08-2.37	0.77 (0.90) 0.02-3.73	1.02 (0.72) 0.02-2.82	1.08 (0.92) 0.09-3.13
MALP-EM	0.89 (0.65) 0.00-2.09	0.59 (0.44) 0.01-1.39	0.74 (0.48) 0.02-1.92	0.92 (0.60) 0.01-1.79
FAST	0.95 (0.63) 0.07-2.10	0.69 (0.48) 0.07-1.90	0.72 (0.56) 0.05-1.87	0.71 (0.64) 0.10-1.93
FreeSurfer	1.64 (1.05) 0.26-4.08	2.72 (2.88) 0.34-13.31	1.95 (2.26) 0.03-7.95	2.62 (2.27) 0.15-7.77

Table 3: Spearman's ranked correlation for segmentations of total GM for all HD participants included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
			PreHD-A			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.777	1				
SPM 12 Segment	0.896	0.666	1			
ANTs Atropos	0.732	0.833	0.737	1		
MALP-EM	0.808	0.950	0.749	0.844	1	
FSL FAST	0.741	0.785	0.770	0.926	0.874	1
FreeSurfer	0.865	0.959	0.752	0.805	0.952	0.802
			PreHD-B			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.770	1				
SPM 12 Segment	0.932	0.889	1			
ANTs Atropos	0.726	0.934	0.851	1		
MALP-EM	0.842	0.964	0.941	0.919	1	
FSL FAST	0.883	0.848	0.917	0.857	0.878	1
FreeSurfer	0.871	0.925	0.952	0.838	0.938	0.887
			HD1			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.737	1				
SPM 12 Segment	0.857	0.672	1			
ANTs Atropos	0.847	0.818	0.779	1		
MALP-EM	0.793	0.955	0.702	0.886	1	
FSL FAST	0.845	0.731	0.815	0.917	0.841	1
FreeSurfer	0.884	0.857	0.875	0.829	0.895	0.826
			HD2			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.441	1				
SPM 12 Segment	0.758	0.555	1			
ANTs Atropos	0.644	0.820	0.671	1		
MALP-EM	0.633	0.917	0.719	0.860	1	
FSL FAST	0.827	0.738	0.768	0.908	0.853	1
FreeSurfer	0.605	0.802	0.796	0.681	0.883	0.726

Table 4: Spearman's ranked correlation for segmentations of cortical GM for all HD participants included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
			PreHD-A			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.815	1				
SPM 12 Segment	0.883	0.722	1			
ANTs Atropos	0.865	0.901	0.844	1		
MALP-EM	0.836	0.970	0.755	0.886	1	
FSL FAST	0.848	0.848	0.847	0.920	0.874	1
FreeSurfer	0.868	0.955	0.811	0.917	0.973	0.923
			PreHD-B			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.747	1				
SPM 12 Segment	0.926	0.863	1			
ANTs Atropos	0.762	0.943	0.896	1		
MALP-EM	0.830	0.974	0.922	0.947	1	
FSL FAST	0.818	0.934	0.925	0.916	0.955	1
FreeSurfer	0.887	0.836	0.935	0.848	0.874	0.925
			HD1			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.734	1				
SPM 12 Segment	0.868	0.716	1			
ANTs Atropos	0.869	0.898	0.844	1		
MALP-EM	0.735	0.973	0.711	0.914	1	
FSL FAST	0.820	0.764	0.862	0.902	0.782	1
FreeSurfer	0.820	0.893	0.898	0.910	0.901	0.854
			HD2			
SPM 8 Unified Segment	1					
SPM 8 New Segment	0.411	1				
SPM 12 Segment	0.826	0.576	1			
ANTs Atropos	0.692	0.814	0.741	1		
MALP-EM	0.638	0.910	0.749	0.869	1	
FSL FAST	0.803	0.699	0.835	0.919	0.860	1
FreeSurfer	0.729	0.729	0.887	0.832	0.868	0.880

Table 5: Spearman's ranked correlation for segmentations of frontal GM for all HD groups included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
Controls						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.807	1.000				
SPM 12 Segment	0.918	0.935	1.000			
ANTs Atropos	0.800	0.972	0.951	1.000		
MALP-EM	0.860	0.975	0.953	0.954	1.000	
FSL FAST	0.891	0.879	0.940	0.907	0.925	1.000
FreeSurfer	0.863	0.940	0.914	0.900	0.965	0.902
PreHD-A						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.811	1.000				
SPM 12 Segment	0.919	0.749	1.000			
ANTs Atropos	0.839	0.865	0.859	1.000		
MALP-EM	0.832	0.962	0.737	0.841	1.000	
FSL FAST	0.820	0.638	0.802	0.776	0.692	1.000
FreeSurfer	0.835	0.925	0.758	0.878	0.943	0.788
PreHD-B						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.761	1.000				
SPM 12 Segment	0.902	0.926	1.000			
ANTs Atropos	0.770	0.971	0.935	1.000		
MALP-EM	0.821	0.968	0.949	0.943	1.000	
FSL FAST	0.856	0.830	0.892	0.824	0.847	1.000
FreeSurfer	0.826	0.950	0.949	0.932	0.958	0.917
HD1						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.806	1.000				
SPM 12 Segment	0.818	0.689	1.000			
ANTs Atropos	0.851	0.914	0.803	1.000		
MALP-EM	0.761	0.976	0.630	0.908	1.000	
FSL FAST	0.678	0.713	0.677	0.803	0.713	1.000
FreeSurfer	0.875	0.896	0.868	0.925	0.875	0.762
HD2						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.582	1.000				
SPM 12 Segment	0.869	0.704	1.000			
ANTs Atropos	0.738	0.805	0.791	1.000		
MALP-EM	0.662	0.970	0.777	0.881	1.000	
FSL FAST	0.839	0.681	0.788	0.877	0.784	1.000
FreeSurfer	0.738	0.889	0.893	0.818	0.916	0.777

Table 6: Spearman's ranked correlation for segmentations of temporal GM for all HD participants included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
			Controls			
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.863	1.000				
SPM 12 Segment	0.972	0.932	1.000			
ANTs Atropos	0.861	0.972	0.930	1.000		
MALP-EM	0.874	0.967	0.949	0.981	1.000	
FSL FAST	0.928	0.947	0.954	0.944	0.937	1.000
FreeSurfer	0.918	0.956	0.951	0.956	0.939	0.975
			PreHD-A			
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.853	1.000				
SPM 12 Segment	0.865	0.826	1.000			
ANTs Atropos	0.908	0.944	0.856	1.000		
MALP-EM	0.844	0.937	0.818	0.964	1.000	
FSL FAST	0.914	0.908	0.890	0.970	0.943	1.000
FreeSurfer	0.904	0.949	0.878	0.952	0.944	0.946
			PreHD-B			
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.767	1.000				
SPM 12 Segment	0.941	0.901	1.000			
ANTs Atropos	0.800	0.990	0.926	1.000		
MALP-EM	0.824	0.971	0.934	0.970	1.000	
FSL FAST	0.908	0.895	0.962	0.899	0.925	1.000
FreeSurfer	0.896	0.896	0.956	0.890	0.916	0.974
			HD1			
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.768	1.000				
SPM 12 Segment	0.892	0.689	1.000			
ANTs Atropos	0.842	0.868	0.874	1.000		
MALP-EM	0.830	0.943	0.737	0.895	1.000	
FSL FAST	0.820	0.773	0.785	0.821	0.863	1.000
FreeSurfer	0.821	0.723	0.889	0.833	0.809	0.874
			HD2			
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.641	1.000				
SPM 12 Segment	0.886	0.749	1.000			
ANTs Atropos	0.779	0.829	0.839	1.000		
MALP-EM	0.770	0.925	0.865	0.935	1.000	
FSL FAST	0.794	0.827	0.922	0.844	0.863	1.000
FreeSurfer	0.746	0.893	0.907	0.866	0.926	0.926

Table 7: Spearman's ranked correlation for segmentations of parietal GM for all HD participants included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
Controls						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.791	1.000				
SPM 12 Segment	0.925	0.879	1.000			
ANTs Atropos	0.854	0.947	0.942	1.000		
MALP-EM	0.833	0.981	0.877	0.951	1.000	
FSL FAST	0.846	0.937	0.905	0.912	0.932	1.000
FreeSurfer	0.898	0.937	0.900	0.928	0.967	0.935
PreHD-A						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.845	1.000				
SPM 12 Segment	0.908	0.764	1.000			
ANTs Atropos	0.854	0.884	0.847	1.000		
MALP-EM	0.812	0.959	0.741	0.871	1.000	
FSL FAST	0.869	0.892	0.845	0.964	0.862	1.000
FreeSurfer	0.878	0.944	0.830	0.944	0.928	0.971
PreHD-B						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.731	1.000				
SPM 12 Segment	0.920	0.863	1.000			
ANTs Atropos	0.785	0.941	0.931	1.000		
MALP-EM	0.856	0.956	0.940	0.943	1.000	
FSL FAST	0.866	0.863	0.952	0.899	0.914	1.000
FreeSurfer	0.800	0.926	0.922	0.931	0.947	0.961
HD1						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.762	1.000				
SPM 12 Segment	0.899	0.705	1.000			
ANTs Atropos	0.901	0.916	0.863	1.000		
MALP-EM	0.787	0.968	0.761	0.944	1.000	
FSL FAST	0.908	0.853	0.881	0.937	0.881	1.000
FreeSurfer	0.904	0.884	0.902	0.949	0.925	0.970
HD2						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.587	1.000				
SPM 12 Segment	0.836	0.768	1.000			
ANTs Atropos	0.681	0.928	0.815	1.000		
MALP-EM	0.632	0.959	0.777	0.934	1.000	
FSL FAST	0.744	0.947	0.899	0.925	0.938	1.000
FreeSurfer	0.692	0.869	0.883	0.896	0.869	0.938

Table 8: Spearman's ranked correlation for segmentations of occipital GM for all HD participants included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
Controls						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.712	1.000				
SPM 12 Segment	0.907	0.854	1.000			
ANTs Atropos	0.761	0.930	0.914	1.000		
MALP-EM	0.826	0.956	0.898	0.928	1.000	
FSL FAST	0.877	0.879	0.921	0.911	0.942	1.000
FreeSurfer	0.863	0.895	0.875	0.883	0.939	0.914
PreHD-A						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.871	1.000				
SPM 12 Segment	0.886	0.770	1.000			
ANTs Atropos	0.893	0.859	0.793	1.000		
MALP-EM	0.859	0.940	0.794	0.887	1.000	
FSL FAST	0.908	0.856	0.793	0.958	0.902	1.000
FreeSurfer	0.871	0.908	0.794	0.883	0.946	0.910
PreHD-B						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.553	1.000				
SPM 12 Segment	0.848	0.814	1.000			
ANTs Atropos	0.639	0.899	0.850	1.000		
MALP-EM	0.753	0.902	0.911	0.857	1.000	
FSL FAST	0.815	0.737	0.902	0.812	0.848	1.000
FreeSurfer	0.657	0.774	0.868	0.762	0.862	0.764
HD1						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.746	1.000				
SPM 12 Segment	0.958	0.741	1.000			
ANTs Atropos	0.899	0.851	0.917	1.000		
MALP-EM	0.844	0.956	0.844	0.905	1.000	
FSL FAST	0.976	0.788	0.967	0.935	0.881	1.000
FreeSurfer	0.922	0.787	0.958	0.922	0.878	0.949
HD2						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.665	1.000				
SPM 12 Segment	0.931	0.794	1.000			
ANTs Atropos	0.755	0.892	0.841	1.000		
MALP-EM	0.785	0.959	0.877	0.850	1.000	
FSL FAST	0.896	0.865	0.932	0.919	0.899	1.000
FreeSurfer	0.809	0.832	0.890	0.800	0.898	0.896

Table 9: Spearman's ranked correlation for segmentations of insula GM for all HD participants included in the current study.

	SPM 8 Unified Segment	SPM 8 New Segment	SPM 12 Segment	ANTs Atropos	MALP-EM	FSL FAST
Controls						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.830	1.000				
SPM 12 Segment	0.925	0.930	1.000			
ANTs Atropos	0.818	0.961	0.874	1.000		
MALP-EM	0.704	0.791	0.709	0.811	1.000	
FSL FAST	0.839	0.939	0.868	0.953	0.837	1.000
FreeSurfer	0.832	0.883	0.819	0.890	0.884	0.925
PreHD-A						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.922	1.000				
SPM 12 Segment	0.928	0.853	1.000			
ANTs Atropos	0.950	0.965	0.902	1.000		
MALP-EM	0.866	0.944	0.731	0.920	1.000	
FSL FAST	0.937	0.914	0.893	0.971	0.863	1.000
FreeSurfer	0.919	0.947	0.848	0.965	0.907	0.968
PreHD-B						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.922	1.000				
SPM 12 Segment	0.928	0.853	1.000			
ANTs Atropos	0.950	0.965	0.902	1.000		
MALP-EM	0.866	0.944	0.731	0.920	1.000	
FSL FAST	0.937	0.914	0.893	0.971	0.863	1.000
FreeSurfer	0.919	0.947	0.848	0.965	0.907	0.968
HD1						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.878	1.000				
SPM 12 Segment	0.916	0.937	1.000			
ANTs Atropos	0.908	0.965	0.973	1.000		
MALP-EM	0.874	0.881	0.776	0.815	1.000	
FSL FAST	0.938	0.958	0.979	0.982	0.851	1.000
FreeSurfer	0.887	0.929	0.928	0.928	0.866	0.959
HD2						
SPM 8 Unified Segment	1.000					
SPM 8 New Segment	0.746	1.000				
SPM 12 Segment	0.829	0.790	1.000			
ANTs Atropos	0.826	0.850	0.917	1.000		
MALP-EM	0.753	0.872	0.777	0.808	1.000	
FSL FAST	0.928	0.857	0.878	0.904	0.851	1.000
FreeSurfer	0.811	0.979	0.818	0.869	0.869	0.908

10. PUBLICATIONS

Chapter 3: GM methods comparison

This paper was published on the analysis conducted in chapter 3.

Johnson, E.B., Gregory, S., Johnson, H.J., Durr, A., Leavitt, B.R., Roos, R.A., Rees, G., Tabrizi, S.J., Scahill, R.I., 2017. Recommendations for the Use of Automated Gray Matter Segmentation Tools: Evidence from Huntington's Disease. *Frontiers in Neurology.* 8, 519. doi:10.3389/fneur.2017.00519