Exploring clinical applications for a novel multi-task functional assessment: matching appropriate technology to clinical need

This thesis is submitted for the degree of Doctor of Philosophy (Engineering) at Cardiff University

Samuel Dominic Woodgate

Supervisors:

Professor Cathy Holt

Professor Monica Busse-Morris

Dr Yulia Hicks

Professor Anne Rosser

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

June 2021

Summary

Huntington's Disease (HD) is an autosomal-dominant, progressive, and ultimately fatal neurodegenerative disorder for which no proven disease-modifying therapies currently exist. Potential therapies are under investigation however their development relies on sensitive clinical assessments of which there is a recognised deficit in HD. This deficit has led to an ongoing push to develop a new wave of clinical assessments for HD. One such assessment is the newly developed Clinch Token Transfer Test (C3t), a multi-task assessment which has been shown to be sensitive to the motor, cognitive, and functional deficits seen in HD. This thesis concerns the continued development of the C3t as a modern clinical assessment.

In chapter 2 it is shown that the C3t protocol can be simplified significantly with many measures found to be redundant. The measures recommended for retention are then shown to be suitable for accurately estimating modern composite measures of disease state along with general motor function. Additionally, it is found that whilst the C3t is sensitive to general motor function it appears to be insensitive to chorea, a common early-stage motor symptom seen in HD. Chapter 3 details the development of a data collection platform built for the C3t which enables large-scale multi-site data collection and the fusion of C3t data with sensor data. The development process and performance of this system is presented as a case study and used to suggest recommendations for future similar work. Finally, chapter 4 develops the C3t into an instrumented assessment using wrist-worn accelerometers. Data collected from the instrumented C3t are used to develop a series of features which are shown to be suitable for accurately estimating whole-body and upper-body chorea. Multiple variations of these features are generated to explore the impact different aspects of the features have on their relationship with chorea to help guide future work.

Contents

Acknowledgements

First and foremost, Philippa, I won't do the whole "words can't express" thing because they very much can – throughout my entire PhD you have been a consistent source of advice, encouragement, and companionship. You're the sort of person who constantly pulls up and supports everyone else, a quality the importance of which I think cannot be overstated to have in a team. Our talks have not only helped my academic work, but they have helped me progress personally and professionally from the very green nervous computer science student that joined all those years ago. Thank you.

My supervisors, Cathy, Monica, Yulia, and Anne. The advice and guidance each of you have given me over the course of my studies has naturally been invaluable. Cathy, you helped me develop a true engineering mindset rather than someone who just knows how to code. As I transition from academia into industry the lessons you taught me about communication and considered design have already served me well. Monica, without you teaching me how to work with patients and clinicians the work presented in this thesis would not have been possible, thank you for your constant guidance and for being so patient with me. Yulia, as my supervisor whose background is the closest to my own having someone to directly relate to was always very welcome. Anne, I always found your presence in clinic extremely calming, your help with collecting and understanding data from patients was critical to this project.

The staff and patients at the Haydn Ellis Huntington's Clinic and the CTR, without which the work presented in this thesis would not have been possible. Too many to name, but each one of you contributed in a meaningful way to this work, thank you.

Nidal, our late nights in the lab and later nights in the city where I think what we both needed. You've been a true friend over the last few years, I count myself very lucky to have met you.

My partner Caitlin and our friends Will, Kate, Sam, and Johanna (a.k.a., the Quasi-useless Club), I won't pretend any of us are our most productive when we're together, but the long lunches and late dinners certainly took the edge off the times when we were. s

My parents Andrew and Mary, for all your love and support, too many things to name individually, but you really are the best parents anyone could wish for, as I hope you know. Similarly, my maternal grandmother, June Brailsford, you've always been incredibly supportive over the years, I only wish grandpa could have been around to discuss things with.

Last but not least, June and Chris, our neighbours who during the first summer of the Covid-19 pandemic gave me a place to stay whilst my parents were sheltering. Thank you for being the true definition of good neighbours.

For my maternal grandfather, Edward Noel Brailsford.

Abbreviations

ADL	Activities of daily living
BAT	Baseline Alphabet Task
BTT	Baseline Transfer Task
BVT	Baseline Value Task
C3t	Clinch Token Transfer Test
CAG	Cytosine, adenine, guanine repeats
CRF	Case report form
CRUD	Create, read, update, delete
CTT	Complex Transfer Task
CTTI	Clinical Trials Transformation Initiative
CUHDRS	Composite Unified Huntington's Disease Rating Scale
CVT	Complex Value Task
DBS	Disease Burden Score
DCL	Diagnostic Confidence Level
DTT	Dual Transfer Task
eCRF	Electronic Case Report Form
EMG	Electromyography
FAS	Functional Assessment
HD	Huntington's Disease
HTTPS	Hypertext Transfer Protocol Secure
IMU	Inertial Measurement Unit
IS	Independence Scale
LFT	Letter Fluency Test
MAE	Mean Absolute Error
MBT	Money Box Test
MULMS	Modified Upper-limb Motor Score
N-MAE	Normalised Mean Absolute Error
OOP	Object Oriented Programming
PBA	Problem Behaviours Assessment
PBA-s	Problem Behaviours Assessment Short
PD	Parkinson's Disease

PIHD	Prognostic Index for HD
PINHD	Prognostic Index Normalised for HD
PK	Primary Key
RDCP	Remote Data Collection Platform
RMSE	Root Mean Squared Error
SCWT	Stroop Colour Word Test
SDMT	Symbol Digit Modalities Test
SNTP	Simple Network Time Protocol
SWRT	Stroop Word Reading Test
TFC	Total Functional Capacity
TLS	Transport Layer Security
TTT	Triple Transfer Task
UHDRS	Unified Huntington's Disease Rating Scale
UHDRS-b	Unified Huntington's Disease Rating Scale Behaviour
UHDRS-TMS	Unified Huntington's Disease Rating Scale Total Motor Score
UTC	Universal Time Coordinated

Abstract

Huntington's Disease (HD) is an autosomal-dominant, progressive, and ultimately fatal neurodegenerative disorder which results in complex array of motor, cognitive, behavioural, and functional symptoms. At present no disease modifying therapies are available however numerous potential therapies are under active development. Due to the progressive nature of the disease, there is a particular focus on therapies that target the earliest stages of HD, with a view to slowing progression before too much damage has occurred.

To help prove the efficacy of such potential therapies, as well as for facilitating effective clinical management, sensitive assessments of HD disease state are required. It has however been repeatedly shown throughout the literature that existing assessments methods used in HD are unsuitable for measuring subtle changes in disease symptom progression. This presents a clear problem for the development of potential treatments as well as clinical management and as such there is an ongoing drive to develop new assessment strategies for HD.

In response to this need for new HD assessment strategies (specifically regarding functional symptoms) the Clinch Token Transfer Test (C3t), a timed upper-body dexterity test, was developed. The C3t has been shown in previous work to be sensitive to various gold-standard HD assessments and an instrumented variant shown to related to general upper-body motor function. This thesis expands on this previous work with the goal simplifying uptake of the C3t, providing further evidence of the C3ts utility in HD assessment and exploring its relationship with chorea, a common early-stage HD motor symptom, using data from wrist-mounted accelerometers worn during the test. Additionally, this thesis details and critiques the development and deployment of a remote data collection platform (RDCP) designed for the C3t which facilitated the collection of much of the data used in this study.

First, understanding of the C3t and the scores it contains was developed using C3t and clinical data from one-hundred and five HD gene-positive participants of varying disease stages (pre-manifest to TFC Stage 3) of which thirty-three had 1-month and 12-month follow-up data. Four clinical measures were included in the study – the UHDRS-TMS, the Composite Unified Huntington's Disease Rating Scale (CUHDRS), the Prognostic Index Normalised for HD (PIN_{HD}), and a chorea score from the summed chorea components of the UHDRS-TMS. C3t scores were available for all visits, clinical measures were only available for the baseline and 12-month visits. Analysis of the C3t scores distribution within the cohort showed six of the fourteen scores were mostly invariant and so could be removed from the C3t protocol and further analysis. Six additional scores were also ultimately recommended for removal –

two as they were solely derived from the invariant scores, two which showed no relationship with any of the studied clinical measures, and two which extremely high correlations with the C3t time scores but require extra work to produce making them effectively redundant. The two remaining C3t scores, both time-taken scores, were highly correlated with each clinical measure (Spearman's R, UHDRS-TMS r=0.69; CUHDRS r=-0.69; PIN_{HD} r=0.83) and could be used as independent variables in regression models to estimate the CUHDRS and UHDRS-TMS with a low degree of error (normalised mean absolute error (N-MAE), UHDRS-TMS=9.4%; CUHDRS=11.0%). No relationship was found between any C3t score and the summed chorea score. Effect sizes calculated for the C3t scores and each clinical measure between the baseline and 1-month visits (C3t scores only) and baseline and 12-month visits (C3t scores and clinical measures) were inconclusive. Finally, it was found that study site and test version did not impact regression models produced using the C3t time scores to estimate the clinical measures.

As no relationship was observed between any non-instrumented C3t score and the summed chorea score signal features thought to be sensitive to chorea were decided upon and extracted from instrumented C3t data. Data were drawn from fifty-five HD gene-positive participants who wore two GeneActiv tri-axis accelerometers, one on each wrist, whilst taking the C3t. In keeping with recommendations from reviewed literature and expert clinician advice, features were chosen whose hypothesised relationship with chorea would be simple to explain clinically. Two time-domain features were ultimately generated – the number of peaks in a signal and the width between the peaks. To study the impact of different methods of feature generation variations of these features were produced. Variations included generating the features from acceleration and jerk signals, using different high-pass filters prior to feature generation, and combining features generated from different mixes of axes and C3t tasks. Strong correlations were found between the generated features and whole-body chorea (r=0.81), upper-body chorea (r=0.79), and the UHDRS-TMS (r=0.85). These features could also estimate each clinical measure with a low degree of error (N-MAE = 15.3%, 14.8%, and 12.2% respectively). Filter frequency had a large impact on feature quality, with the best performing feature using a bandpass filter of 7.5Hz-0.3Hz, suggesting this may be a good frequency band to use for generating features sensitive to chorea. Axes and task makeup had minimal impact on feature quality. Features generated from jerk tended to outperform those generated from acceleration, however the difference was marginal.

Both sets of analysis relied heavily on data collected using the developed RDCP. The developed system facilitated the synchronisation of timestamps between the C3t task times and sensor recordings. It also facilitated the transmission of C3t data from remote study sites. Although the system was by-large successful several design flaws along with issues involving the GeneActiv accelerometers reduced the amount of data ultimately available. By assessing the issues encountered by study sites six recommendations for future similar research and software platforms were developed.

First, sensors should be chosen based on both technical suitability and usability. In this project sensors were chosen primarily based on technical suitability and availability. In practice however usability of the sensors was found to be poor, with many sites and clinicians reporting significant issues properly using the sensors. Future projects should trial sensors with the clinicians who are using them prior to be selected.

Second, development of the RDCP took place whilst the C3t was still being developed. As such part the way through the platform's development the second version of the C3t was released. This necessitated re-working some of the underlying software, increasing development time. Whilst this can be unavoidable, future work may wish to properly take subsequent test versions into account when designing software systems and building them in a more generic manner such that modifications are as simple as possible to make.

Third, the Waterfall software development methodology was used in this study despite Agile being the methodology typically preferred in industry. The rationale was that academic projects typically have their requirements set far ahead of the project starting and, in such cases, Waterfall can provide a quicker more streamlined development cycle. However, the requirements of the software changed throughout the project making waterfall unsuitable for use. Regardless, future projects should still consider Waterfall as a viable methodology when development software systems for research projects in cases where those projects are fully defined before they are started.

Fourth, whilst training was given to clinicians using the RDCP testing of that training was not conducted. As such although the system appeared straightforward to operate and was found easy to use by local clinical teams, some other teams found the system hard to operate. Future work should conduct at least some testing of any training provided and provide access to an online 'how to use' resource.

Fifth, projects which include data collected from sensor devices, particularly those that include multiple study sites, should implement automatic monitoring of data quality. In this study sensor data was not reviewed until after it had been fully collected. As such, data quality issues were not detected until it was too late to do anything about. Additionally, reviewing sensor data quality is a specialist operation most study managers will have little to no expertise in. As such, engineers working on such projects should develop systems to review data automatically and send reports to specialist personal capable of reviewing sensor data as it comes in.

Finally, any clinician-facing software systems should include a function for reporting issues. In addition to enhanced training clinicians should be able to send reports from within the software itself when aspects of it are not functioning correctly. This would allow the development team to isolate ‘pain points’ within the software and apply fixes proactively.

Chapter 1: Introduction

1.1 Thesis Overview

This thesis details the continued development of the Clinch Token Transfer Task (C3t), a novel clinical research assessment tool designed for using in Huntington's Disease (HD). It should be noted that, as at time of writing the C3t is still in under development. As such, the C3t is currently only suitable as a clinical research assessment tool rather than a tool to inform clinical decision making. However, for brevity the C3t will be referred to as a clinical assessment throughout this thesis.

HD is a devastating, ultimately fatal, autosomal-dominant disease which results in a wide array of functional, behavioural, cognitive, and motor impairments. Due to recognised limitations in current symptom and disease assessment strategies there has been an ongoing drive develop new clinical assessments for HD for a number of years. The C3t is one such assessment; developed originally for assessing functional impairment the C3t has been shown in previous work to be sensitive to several gold-standard assessments of cognitive, motor, and functional impairment. Additionally, measures produced by an instrumented version of the C3t have been shown to be related to general upper-body motor impairment.

Although the C3t is a promising assessment, further work is needed to fully understand and develop its place as a clinical assessment for HD. The overarching aim of the work presented in this thesis was to conduct such further work. More specifically, this thesis details the results of three development directions of the C3t. First, the relationship of scores produced by the C3t with the gold-standard measure of HD motor function and two composite measures of disease state is explored. During this process the C3t scores are critically assessed with a view to removing any redundant scores in order to streamline the test protocol and so aid clinical adoption. Second, a selection of features were developed and extracted from accelerometers worn during the C3t, and their relationship with chorea (a common early-stage HD motor symptom in need of more sophisticated assessment) was assessed. Third, a software system was developed to enable the remote collection of C3t data and to facilitate the collection of instrumented C3t sensor data. This system was then critically analysed to provide recommendations for future similar work, due to the increased importance multi-site remote data collection is playing in clinical studies.

The results of this thesis are contained in chapters 2, 3, and 4 which cover these three distinct but related segments of work.

Chapter Two focuses on the refinement of the C3t protocol along with the confirmation and further development of previous findings using enhanced statistical analysis, predictive machine learning models, and additional datasets. Numerous redundant scores are detected and suggested for removal, significantly simplifying the C3t protocol as well as showing a clear path for further development of the assessment. The results of previous work are confirmed and the understanding of the C3t as a clinical assessment deepened particularly regarding the C3t metrics relationship with recently developed composite assessments.

Chapter Three focuses on the design, development, and deployment of a custom-built remote data collection platform (RDCP) to better facilitate the large-scale collection of C3t & sensor data across disparate study sites. The construction and deployment of such a software platform is a non-trivial but necessary aspect of modern-day clinical research, particularly when sensor-based datasets are being collected. The case-study presented in chapter three showcases the various design considerations, pitfalls, and pain-points of such systems and makes recommendations pertinent to researchers who require similar software platforms for their own studies.

Chapter Four focuses on analysing accelerometery data collected from wrist-mounted accelerometers worn whilst taking the C3t with the goal of developing features which show a relationship to gold-standard measures of chorea. Signal processing and machine learning techniques are utilised to produce a set of features that were hypothesised to be linked to chorea. Additionally, the nature of chorea and expert clinical advice was considered to produce features which are simple to give clinical meaning to. Evidence is provided that a subset of these features is suitable for estimating chorea in HD with a reasonable degree of accuracy, outperforming previous work in the area.

1.2 Chapter Overview

Huntington’s Disease (HD) is a rare neurodegenerative disorder which results in a complex array of debilitating conditions including motor, cognitive, behavioural, and functional deficits.

However, whilst there are numerous clinical assessments used to evaluate HD symptoms there a relatively few designed specifically for use in HD (Clinch, 2017a). Clinch (2017) noticed that in particular there was a lack of objective assessments functional deficits. As such, the Clinch Token Transfer Test (C3t), is a novel, multi-stage token transfer assessment, was developed with the goal of providing a simple, objective measure of function.

The C3t centres around participants transferring a series of tokens as quickly as possible and in the correct order from their starting position on a board into a slotted box. Participants pick up tokens one a time using their non-dominant hand, transfer the token into their dominant hand and then into the box. There are 3 such transfer tasks all of which are slightly different but with the underlying movement being the same. The time taken to complete each task is recorded as the primary measurement. Despite originally being designed to be sensitive to functional deficits, previous work has found the C3t to be linked to many of the symptom domains seen in HD (Bennasar *et al.*, 2016; Clinch *et al.*, 2018). Notably, Bennasar *et al.*, (2016) found that accelerometers worn during the C3t could be used to estimate the severity of general upper-body motor symptoms. Motor symptom assessment is particularly important in HD as is discussed in section 1.3.3. A complete description of the C3t is given in section 1.5.

In summary, although the C3t appears promising as an HD-specific assessment the test is still in its infancy. Thus, this thesis is primarily concerned with the continued development of the C3t and its augmentation via wearable technology to create an instrumented assessment suitable for assessing motor symptoms in Huntington's Disease (HD). Additionally, this thesis seeks to ease adoption of the C3t in clinical research settings. This is accomplished by simplifying the test protocol via an extended analysis of each of its components and creating a digital data collection platform allowing data to be easily collected from multiple disparate study sites.

This first chapter serves to provide an overview of HD, the rationale for why an instrumented assessment of motor symptoms in is required, and the design considerations for creating one.

There are four parts to this chapter.

Part 1 introduces HD, covering its history, symptoms, current gold-standard assessment strategies, their limitations with respect to assessing disease progression, and the role instrumented assessment have to play in the future of HD clinical assessment.

Part 2 reviews alternatives to the current gold-standard assessment strategies for HD motor symptoms. Both traditional clinical assessments and instrumented assessments are covered, the distinction being the latter incorporates digital sensor technology giving access to different types of data (e.g., acceleration). As there is a paucity of instrumented assessments in HD the rationale for their application in HD is partially based on their use in Parkinson's Disease (PD) (it being similar), and, therefore, instrumented assessments in PD are also reviewed.

Part 3 covers the various design considerations of an instrumented assessment suitable for motor disorders in HD and also the requirements for an associated remote data collection platform (RDCP) to allow such data to be collected at scale. For the instrumented assessment, design considerations include the aims of such an assessment, the qualities of the base assessment (clinical or functional), that is to be instrumented, which sensors are appropriate given the aims & base assessment, how relevant data from those sensors can be correctly extracted and finally how the validity of the assessment can be quantified and thus the aims ultimately realised. For the RDCP, design considerations include the required functionality, technical design decisions and principles which should followed.

Part 4, based on the information given in Parts 1 to 3, states the overall objectives of this thesis, and the specific rationales behind them.

1.3 Part 1: Huntington's Disease

1.3.1 Huntington's Disease: Overview

In 1872 Dr George Huntington M.D. published a paper in the Medical and Surgical Reporter of Philadelphia titled "On Chorea" discussing various the aspects of chorea and recommendations for its management (Huntington, 1967). At the end of the paper, he draws attention to what he terms as "hereditary chorea" describing it as being

> "confined to certain and fortunately a few families […] an heirloom from generations away back in the dim past […] spoken of by those in whose veins the seeds of the disease are known to exists, with a kind of horror" (Huntington, 1967).

He then goes on to note core characteristics of the disease, identifying an inheritance pattern which requires only one parent to have been affected, noting its progressive ultimately fatal motor, cognitive and behavioural impairments and estimating the typical age of onset as being around 30 to 40 years.

George Huntington's description of hereditary chorea, now known as Huntington's Disease, bears close resemblance to its modern-day description although with advances in genetics, neurology, and imaging techniques, our understanding is now more complete.

HD is a rare, autosomal dominant, progressive and ultimately fatal neurodegenerative disorder affecting approximately 6-13 people per 100,000 in the general population (Rawlins *et al.*, 2016). It is caused by an expansion of cytosine, adenine, guanine (CAG) polyglutamine repeats from less than 26 in unaffected populations to over 36 in affected populations (Myers, 2004). The effect of different expansions is shown in Table 1. This expansion leads to a mutant version of the Huntingtin protein being developed that causes damage to medium spiny neurons which are found in abundance in the striatum. This damage to the striatum then results in a break in the circuitry of the basal ganglia, a region of the brain associated with amongst other things voluntary motor control, cognition, and emotion. Eventually the damage to the striatum, as well as other sections of the brain, leads to the manifestations of the motor, cognitive and behavioural abnormalities seen in HD (Clinch, 2017a).

Table 1: Relationship of different cytosine, adenine, guanine (CAG) repeats on development of Huntington's Disease. Adapted from Myers, (2004)*.*

CAG Repeat Length	Result on carrier
<26	Normal range
27-35	Unaffected by HD, fathers may transmit a repeat to descendants high enough to cause HD
36-39	Reduced penetrance of HD – some will develop the disease others will not
40-59	Full penetrance, all carries will develop HD
60+	Full penetrance, increased risk of developing Juvenile HD (onset of HD at or less than 20 years of age)

Historically HD was said to present in two distinct stages – the pre-manifest stage during which no symptoms are present, and the manifest stage when symptoms are present. Due to the prevalence of motor symptoms in HD, the distinction of a pre-manifest patient from a manifest one is the presence of motor symptoms which cannot be attributed to anything other than HD (Wild and Tabrizi, 2014). This concept however leads to the erroneous assumption that patients will one-day awake with motor symptoms suddenly present, or that they will abruptly and rapidly deteriorate over just a few days. The truth however is that HD is a lifelong disease that is biologically present from birth, with symptoms progressively worsening over the course of many years. This acceptance of HD as a long-term, slowly progressive disease has led to the common understanding of a third stage, prodromal, which denotes the period during which motor symptoms are beginning to emerge (Wild and Tabrizi, 2014).

A diagnosis of manifest HD, also known as motor onset, typically occurs at around 40 years of age with survival after onset being around 20 years (Myers, 2004; McColgan and Tabrizi, 2018). Onset before 20 years of age is rare (4-10% of all cases) occurring in those with very high CAG repeats (60+) and is distinguished from HD as Juvenile HD (Fusilli *et al.*, 2018).

It is worth noting that whilst a diagnosis of manifest HD still requires the presence of overt motor symptoms, it is known that cognitive and behavioural symptoms can pre-date motor onset by many years (McColgan and Tabrizi, 2018). Due to the subjectively of these symptoms however motor disorders are still typically preferred for diagnostic purposes although this is starting to change (Hersch and Rosas, 2008).

With regard to treatments, at time of writing no disease modifying therapies have been clinically proven, however numerous potential therapies are under active investigation (McColgan and Tabrizi, 2018). A particularly promising avenue of research is in lowering levels of the mutant huntingtin (mHTT) protein, which is responsible for the neurodegeneration seen in HD (Schulte and Littleton, 2011; McColgan and Tabrizi, 2018). Due to the progressive damage seen in HD, therapies which seek to suppress or eliminate the production of the mHTT protein will logically be most beneficial the earlier on during the course of the disease they are applied, ideally in the premanifest phase (Wild and Tabrizi, 2017).

An alternative approach to reducing or stopping the production of mHTT is to repair the damage it causes to the striatum, potentially by way of pluripotent stem cells (Li and Rosser, 2017). Such therapies have been under active investigation for several years, with some potential therapies currently undergoing stage 3 clinical trials (Bachoud-Lévi, Massart and Rosser, 2021). This approach is complimentary to reducing mHTT production, with the former repairing existing damage and the later reducing ongoing damage (Bachoud-Lévi, Massart and Rosser, 2021).

Regardless of whether a proposed therapeutic seeks to repair structural damage or slow disease progression, both approaches will rely on the sensitive assessment of symptoms during large-scale studies to prove their efficacy (Clinch *et al.*, 2018). There are, however, limitations in the assessment methods currently in use.

The capabilities and limitations of current assessments are detailed in section 1.3.3. Broadly speaking however there are three main limitations/criticisms of current assessments – a lack of sensitivity to early-stage symptoms, lack of sensitivity to symptom progression, and subjectivity. Such limitations of existing assessment methods are the primary motivation for this thesis – in short there exists a well-recognised and ongoing need in HD for sensitive assessments suitable for deployment at a large scale.

In order to discuss symptom assessment limitations however first the symptoms themselves must be understood, which is the topic of the following section.

1.3.2 Huntington's Disease: Symptoms

1.3.2.1 The symptom domains of HD

HD symptoms can be said to develop across four domains – cognitive, behavioural, motor, and functional. Each of these domains consist of different symptoms which are dynamic rather than static, evolving over time as the disease progresses (McColgan and Tabrizi, 2018). The list of potential symptoms is extensive and not all patients present with all listed symptoms. Some symptoms, such as chorea, are highly common whilst others, such as obsessive-compulsive disorders, are only experienced by a fraction of patients (McColgan and Tabrizi, 2018). The primary characteristic of HD symptoms is their inter-subject variability, to the extent that two siblings with HD may present with different symptoms, progress at different rates, and thus require different strategies for effective disease management. This variability is part of what makes clinical assessments so challenging and why, as is discussed later, composite assessments are being proposed as part of a comprehensive assessment strategy.

The following subsections give a short overview of each symptom domain with particular attention paid to the motor symptoms as their assessment is critical to this thesis. The assessment of these symptoms and how they can be tracked over time is covered in section 1.3.3.

1.3.2.2 Cognition

A multitude of cognitive changes can occur in people with HD which, similar to other symptoms, progressively worsen over time (Meyer *et al.*, 2012). Common symptoms include psychomotor slowing, attention deficits, decreased executive function and problems with memory, learning & emotion recognition (Craufurd and Snowden, 2014). In the earlier stages of the disease (pre-manifest & early-manifest) lowered processing speed, multi-tasking ability and executive function deficits may be present with the full range of symptoms appearing later as the disease progresses (Papoutsi *et al.*, 2014).

1.3.2.3 Behaviour

Whilst motor and cognitive deficits are arguably the better-known symptoms of HD behavioural abnormalities are also highly prevalent. A large-scale study (n=1766) of the REGISTRY database estimated that up to 87% of people who test gene-positive for HD will display some level of behavioural abnormality (Orth *et al.*, 2011).

A wide range of behavioural symptoms may present in HD including apathy, affective disorders, irritability, mania, psychosis, sexual disorders and suicidal ideation (Craufurd and Snowden, 2014). Apathy is typically the most common disorder seen across all disease stages (~28% prevalence) followed by depression, irritability and obsessive-compulsive behaviours (~13% prevalence) with other disorders being comparatively rare (McColgan and Tabrizi, 2018). It is worth noting that the prevalence of affective disorders (e.g., depression) in HD is thought to be due to some underlying neurological effect of the HD mutation itself rather than as a psychological reaction to the presence of the disease (Craufurd and Snowden, 2014).

1.3.2.4 Motor

Historically motor symptoms have been considered the cardinal symptoms associated with HD. Whilst chorea is by far the best known (to the point that Huntington's Disease used to be known as Huntington's Chorea (Vale and Cardoso, 2015)) numerous other motor symptoms can present during the course of HD including bradykinesia, dysarthria, dysphagia, dystonia, , gait & balance disturbances, oculomotor dysfunction, rigidity and tics (Roos, 2014). Short descriptions of each of these is given in Table 2.

Table 2: Motor symptoms of HD

Motor Symptom	***Description***
Chorea	*Random, sudden, rapid, involuntary, and purposeless movements*
Bradykinesia	*Reduced movement velocity (often progressive in repetitive tasks) and slowness to initiate movement*
Dysarthria	*Slurred / slow speech, difficulty moving tongue or facial muscles*
Dysphagia	*Difficulty swallowing leading to coughing or choking when eating and drinking*
Dystonia	*Repetitive twisting movements, abnormal fixed postures*
Gait & Balance Disturbances	*Abnormal gait & posture, instability whilst walking*
Oculomotor Dysfunction	*Delayed/suppressed saccade initiation and gaze impersistence*
Rigidity	*Stiff and inflexible muscles*
Tics	*Rapid, suppressible movements primarily in face and arms*

Similar to other symptoms seen in HD, the motor symptoms seen are progressive, emerging gradually over the course many years (Roos, 2014). Unlike other symptoms however their progression is relatively more uniform across the population. Throughout the pre-manifest stage motor symptoms are by definition not present although subtle unintended movements may sometimes still be detected by careful clinical observation (Roos, 2014). As the disease progresses into the prodromal stage soft motor symptoms begin to emerge with typical symptoms being small ticks in the extremities, reduced postural stability, saccadic delay and gaze impersistence (Wild and Tabrizi, 2014). Once in the manifest stage there is usually an initial hyperkinetic phase predominantly characterised by chorea although oculomotor dysfunction and tics may also be present. As the disease becomes more advanced hyperkinetic movements begin to plateau before being subsumed by a hypokinetic phase typically dominated by bradykinesia, dystonia and gait & balance disturbances (McColgan and Tabrizi, 2018). In the final stages of HD movement and speech become increasingly restricted as rigidity and dysarthria take hold. Additionally, the presence of dysphagia makes swallowing difficult and ultimately dangerous (Roos, 2014).

1.3.2.5 Functional Capacity

The combination of cognitive, behaviour and motor symptoms seen in HD together result in the decline of patients ability to perform tasks necessary for daily living, termed their 'functional capacity' (Mestre, Busse, *et al.*, 2018). Common deficits are reported in a wide variety of areas with five key areas typically being rated in assessments that seek to capture decline in functional capacity - occupation, handling finances, domestic chores, ability to self-care and genera activities of daily living (Kieburtz *et al.*, 1996).

The progressive nature of cognitive, behavioural, and motor symptoms in HD makes the decline in functional capacity similarly progressive. As a result, decline in functional capacity is a common feature seen throughout manifest HD and its degradation has been shown to be consistent and robust in nature (Meyer *et al.*, 2012). In pre-manifest and prodromal HD its prevalence is less certain with one large-scale study of prodromal participants detecting no decline in functional capacity in 88% of their cohort (Paulsen *et al.*, 2010).

1.3.3 Huntington's Disease: Assessing disease state

1.3.3.1 The Unified Huntington's Disease Rating Scale

Originally developed by the Huntington's Study Group in 1979 and revised in 1999 the Unified Huntington's Disease Rating Scale (UHDRS) is the de-facto standard for them clinical assessment of HD and its symptoms (Kieburtz *et al.*, 1996). Despite being termed a scale, the UHDRS is actually a collection of assessments deemed suitable for assessing the cognitive, behavioural, motor, and functional deficits seen in HD. Assessment using the UHDRS is carried out by trained clinicians with a number of professional bodies providing training and certification for its various component assessments.

Of the four symptom areas covered by the UHDRS, only the behaviour assessment has been supplanted by an alternative measure, namely the Problem Behaviours Assessment (PBA). At time of writing the cognitive, motor, and functional assessments are all still typically considered to be the gold-standard assessments for their respective symptom domains 41 years (21 after revision) after the scale's original development (Winder *et al.*, 2019). This can be evidenced by not only its regular referral throughout the HD literature, but also its application in large-scale observational studies. Possibly the best known of these observational studies is Enroll-HD, a worldwide observational study of Huntington's Disease families, listing the motor, cognitive and functional assessment batteries of the UHDRS as the core components used for assessing their respective symptom domains (Enroll-HD, 2020).

Despite its widespread use, various aspects of the UHDRS assessments have been criticised or questioned in recent years (Reilmann *et al.*, 2011a; Mestre *et al.*, 2016; McColgan and Tabrizi, 2018; Mestre, Bachoud-Lévi, *et al.*, 2018; Mestre, Busse, *et al.*, 2018; Mestre, Forjaz, *et al.*, 2018; Winder *et al.*, 2019). Whilst some of these criticisms have been addressed, such as rigorous training to improve inter-rater reliability, others, such as low sensitivity, are intrinsic characteristics of the assessments and as such cannot be so easily overcome.

The following subsections provide an overview and critique of the UHDRS with a specific focus on detecting temporal changes; staging HD, the four symptom domains, and the more contemporary approach that provides a composite outcome.

1.3.3.2 Staging HD using the UHDRS

HD is typically split into three distinct stages – pre-manifest, prodromal and manifest (Wild and Tabrizi, 2014). A person who tests positive for the HD gene (known as being 'gene-positive') is said to be in the pre-manifest stage of the disease until overt motor symptoms start to manifest. Once soft motor symptoms begin to develop, patients may be said to have entered the prodromal stage. Finally, once a person with the HD gene develops "*the unequivocal presence of an otherwise unexplained extrapyramidal movement disorder*" motor onset is said to have occurred and the manifest stage begins (Wild and Tabrizi, 2014).

Throughout the pre-manifest and prodromal stages, the primary clinical focus is on detecting the presence of motor symptoms. To achieve this, expert clinicians assess patients using the UHDRS motor assessment (commonly referred to as the Total Motor Score (UHDRS-TMS)) and the Diagnostic Confidence Level (DCL). The UHDRS-TMS consists of 31 ordinally-rated assessments of motor function where higher scores indicate increasingly severe symptoms/worse motor task performance. The DCL is a 5-level ordinal rating system which quantifies the confidence level of a clinician that motor symptoms, if present, are due to HD. DCL values range from 0 (no motor abnormalities present) to 4 (≥99% confidence motor abnormalities due to HD are present). The DCL values and their associated confidence requirements are shown in Table 3 and the UHDRS-TMS is covered in more detail in section 1.3.3.5.

Table 3: Diagnostic Confidence Level (adapted from Wild and Tabrizi, (2014)*)*

Diagnostic Confidence Level	***Description***
0	*Normal (no abnormalities)*
1	*Non-specific motor abnormalities (< 50% confidence)*
2	*Motor abnormalities that may be signs of HD (50 -89% confidence)*

3	*Motor abnormalities that are likely signs of HD (90 – 90% confidence)*
4	*Motor abnormalities that are unequivocal signs of HD (≥99% confidence)*

In terms of staging, the specific DCL and UHDRS-TMS requirements for a diagnoses of motor manifestation to be made is a DCL of 4 and UHDRS-TMS > 5. The difference between pre-manifest and prodromal stages however is less rigidly defined. One suggestion has been to define the prodromal stage as the period during which *any* symptoms that may be due to HD begin to emerge (Wild and Tabrizi, 2014). This then necessitates the definition of fourth less widely used stage, 'perimanifest', defined as the period when specifically motor symptoms start showing (i.e., when 0 < DCL < 4) (Wild and Tabrizi, 2014). Typically, throughout the literature however the prodromal stage is used to refer the period during which motor symptoms begin to become detectable.

Once motor manifestation has occurred clinical focus switches from detecting the initial signs of HD motor symptom development to assessing the diseases' impact on a patient's daily life. To this end, the UHDRS Functional Assessment and its resultant summary metric, the Total Functional Capacity score (TFC), are used to assess disease progression during the manifest stage. The UHDRS Functional Assessment is a clinician rated ordinal-scale assessment of the impact HD is having on a patient in terms of various aspects of daily living. The TFC is created by summing the results of UHDRS Functional Assessment. TFC scores range between 13 and 0 with lower values indicating a greater impact and thus suggestive of a more advanced disease stage.

Using ranges of TFC values, Shoulson and Fahn, (1979) subdivided the manifest stage into 5 'TFC Stages' where TFC Stage 1 (TFC range 13-11) encompasses the beginning of the manifest disease and TFC Stage 5 (TFC = 0) refers to its final stages. Table 4 shows the cut-offs for each of the 5 TFC stages. Both the UHDRS Functional Assessment and TFC are discussed more thoroughly in section 1.3.3.6.

Whilst staging HD is vital to proper clinical management, the TFC stages are notably coarse. Previous work, as is discussed at length in section 1.3.3.6, has found they limited potential for noticing small changes in disease state over short periods of time. The assessment and tracking of individual symptoms have been found to be better suited for this use case, as is discussed throughout 1.3.3.

Table 4: Total Functional Capacity disease stages, associated ranges and descriptors (adapted from Wild and Tabrizi, (2014)*)*

TFC Stage	***TFC Range***	***Descriptor***
1	*13-11*	*Early*
2	*10-7*	
3	*6-4*	*Moderate*
4	*3-1*	*Advanced*
5	*0*	

1.3.3.3 Assessing the cognitive domain: The UHDRS Cognitive Assessment

The UHDRS cognitive assessment battery is the most widely used method for assessing cognitive dysfunction in HD and is listed as the core cognitive assessments in the global Enroll-HD study (Enroll-HD, 2020). The battery consists of three tests, the Symbol Digit Modalities Test (SDMT), the Stroop Colour Word Test (SCWT) and the Letter Fluency Test (LFT) (Mestre, Bachoud-Lévi, *et al.*, 2018).

The SDMT, designed to assess psychomotor speed, memory, visual attention and symbolic encoding (Mestre, Bachoud-Lévi, *et al.*, 2018), presents participants with a set of paired symbols and numerals alongside a series of randomly ordered symbols with blank boxes where the numerals should be (Smith, 1968). Participants are given 90 seconds to fill in as many of the blank boxes with the correct numeral as possible. At the end of 90 seconds, the number of correct and incorrect numerals entered is counted to provide the test score.

The SCWT has participants complete three assessments – first read the names of colours printed in blank ink, second state the name of coloured patches of ink and third state the name of the colour used to write the name of a different colour (e.g., the word 'red' might be printed in blue to which participants should respond 'blue') (Stroop, 1935; Scarpina and Tagini, 2017). The SCWT is designed to measure cognitive flexibility, selective attention, response inhibition and psychomotor speed (Mestre, Bachoud-Lévi, *et al.*, 2018). One of the key assessments in the SCWT is the Stroop Word Reading component, referred to here as the Stroop Word Reading Test (SWRT).

The LFT, sometimes also called the Phonemic Fluency Test, has participants state as many words as they can in one minute that start with a given letter (Benton, 1968). The test is conducted three times using a different letter each time with the total number of correct responses being used as the final score and is taken to measure executive function and language skills (Mestre, Bachoud-Lévi, *et al.*, 2018).

Mixed results have been observed when using the SDMT, SCWT/SWRT and LFT as indictors of disease progression. In one series of studies progression was observed over 12-, 24- and 36-month periods in both the SDMT and SWRT for an early-HD cohort (Tabrizi *et al.*, 2011, 2012, 2013). Over a 36-month period both assessments also showed changes in a pre-manifest cohort estimated to be less than 10.8 years from motor onset relative to controls (Tabrizi *et al.*, 2013). This effect was however not seen in another pre-manifest cohort greater than 10.8 years from onset in the same study. In another study which looked at longitudinal changes in early manifest HD, taking 8 samples over 36 months, changes in the SDMT, SCWT and LFT were observed however they were small and erratic (Meyer *et al.*, 2012). Meyer *et al.*, (2012) concluded that for tracking changes in early manifest HD motor and functional measures were much more reliable. The discrepancy between the two studies may be explained by the significantly larger sample size and more regular measurements of Meyer *et al.*, (2012) (early manifest n=379) compared to Tabrizi *et al.*, (2011, 2012, 2013) (early manifest 12-month n=114; 24-month n=116; 36-month n=97).

1.3.3.4 Assessing the behaviour domain: The UHDRS Behaviour Assessment & Problem Behavioural Assessment Short Version

The UHDRS behaviour assessment (UHDRS-b) was the original behavioural assessment used to assess HD (Kieburtz *et al.*, 1996) however the Problem Behaviour Assessment Short Version (PBA-s) has replaced it as the main assessment method for behavioural abnormalities in HD (McColgan and Tabrizi, 2018).

The PBA-s uses a structured interview style to assess the frequency and severity of different behavioural symptoms (Callaghan *et al.*, 2015). Both frequency and severity are rated on a scale of 0-4 with 0 indicating the symptom either rarely or never occurs and 4 indicating the symptom occurs daily or almost daily for all or most of the day and is severe in nature. A total score is calculated for each symptom by multiplying their respective frequency and severity scores. A composite behaviour score may be produced by summing each of these total scores (Callaghan *et al.*, 2015). The areas assessed by the PBA-s along with the rating scales used are shown in Table 5.

Table 5: Problem Behaviour Assessment (Short Version) (PBA-s) assessment areas & rating scales, both severity and frequency measures are reported for each behavioural symptom, adapted from McNally *et al.*, (2015)

Behavioural Symptom	**Severity Measure**
Depressed Mood	0 = not present
Suicidal Ideation	1 = slight, questionable
Anxiety	2 = mild (present, not a problem)
Irritability	3 = moderate (symptom causing problem)
Angry or aggressive behaviour	4 = severe (almost intolerable for carer)
Apathy	**Frequency Measure**
Preservative thinking or behaviour	0 = never/almost never
Obsessive-compulsive behaviour	1 = seldom (less than once per week)
Paranoid thinking or delusions	2 = sometimes (up to 4 times per week)
Hallucinations	3 = frequently (most days or 5, 6, 7 times a week)
Disoriented Behaviour	4 = daily or almost daily for most or all of a day

Although there is a link between functional decline and the existence & severity of behavioural symptoms (specifically apathy, irritability, and depression) they do not appear to be suited to measuring disease progression (McColgan and Tabrizi, 2018). The TRACK-HD study found that over 12-, 24- and 36-month intervals there was a deterioration in the PBA-s measure of apathy in early manifest HD relative to controls at 24 and 36 months only, and that there was an association between the PBA-s composite and functional decline at 36-months (Tabrizi *et al.*, 2011, 2012, 2013). The same series of studies found no detectable change in a pre-manifest cohort for any PBA-s measure including the composite.

It has been suggested that this lack of detectable change is because many of the more common behavioural symptoms (e.g., depression & anxiety) in HD can be controlled with pharmacological intervention (McColgan and Tabrizi, 2018). Current assessments may also of course simply be too insensitive to properly assess symptom progression. This is one of the drivers, along with the devastating impact such symptoms can have on the individual and their carers quality of life, for developing methods to objectively and ideally more sensitively assess such behavioural symptoms (McLauchlan, 2018).

1.3.3.5 Assessing the motor domain: The UHDRS Motor Assessment

The UHDRS motor assessment is the de-facto standard for assessing motor symptoms in HD (McColgan and Tabrizi, 2018). It consists of 15 items with each item containing one or more assessments, for example the gait item is assessed once whilst the maximal chorea item is assessed seven times (face, trunk, mouth, and extremities) (Kieburtz *et al.*, 1996). A total of 31 individual assessments are conducted each of which is rated on a scale of 0 to 4 where higher numbers indicate increased severity/worse performance. The UHDRS-TMS is calculated by summing the scores of each assessment producing a score ranging between 0 to 128. Depending on the cohort being studied the DCL may also be added to the UHDRS-TMS, bringing the maximum score to 132. Table 6 lists each of the UHDRS motor assessments items, assessment areas and how they are rated.

Table 6: UHDRS Motor Assessment items, number of assessments per item and scales used per assessment (adapted from Kieburtz *et al.*, (1996)*)*

Item	**Assessments Areas (n=)**	**Scale**
Ocular Pursuit	Horizontal & vertical (n=2)	0 = complete (normal) 1 = jerky movement 2 = interrupted pursuits/full range 3 = incomplete range 4 = cannot pursue
Saccade Initiation	Horizontal & vertical (n=2)	0 = normal 1 = increased latency only 2 = suppressible blinks or head movements to initiate 3 = un-suppressible head movements 4 = cannot initiate saccades
Saccade Velocity	Horizontal & vertical (n=2)	0 = normal 1 = mild slowing 2 = moderate slowing 3 = severely slow, full range 4 = incomplete range
Dysarthria	N/A (n=1)	0 = normal 1 = unclear, no need to repeat 2 = must repeat to be understood 3 = mostly incomprehensible 4 = mute

Tongue Protrusion	N/A (n=1)	0 = can hold tongue fully protruded for 10 seconds 1 = cannot keep fully protruded for 10 seconds 2 = cannot keep fully protruded for 5 seconds 3 = cannot fully protrude tongue 4 = cannot protrude tongue beyond lips
Maximal Dystonia	Trunk & extremities (n=5)	0 = absent 1 = slight/intermittent 2 = mild/common or moderate/intermittent 3 = moderate/common 4 = marked/prolonged
Maximal Chorea	Face, mouth, trunk & extremities (n=5)	0 = absent 1 = slight/intermittent 2 = mild/common or moderate/intermittent 3 = moderate/common 4 = marked/prolonged
Retropulsion Pull Test	N/A (n=1)	0 = normal 1 = recovers spontaneously 2 = would fall if not caught 3 = tends to fall spontaneously 4 = cannot stand
Finger Taps	Right & left (n=2)	0 = normal (≥ 15/5 sec.) 1 = mild slowing and or reduction in amplitude (11-14/5 sec.) 2 = Moderately impaired. Definite and early fatiguing. May have occasional arrests in movement (7-10/5 sec.). 3 = Severely impaired. Frequent hesitation in initiating movements or arrests in ongoing movements (345 sec.) 4 = Can barely perform the task (0-2/5 sec.)
Pronate/Supinate Hands	Right & left (n=2)	0 = normal 1 = mild slowing and/or irregular 2 = moderate slowing and irregular 3 = severe slowing and irregular 4 = cannot perform

Luria	N/A (n=1)	0 = 34 in 10 seconds, no cue 1 = <4 in 10 seconds, no cue 2 = 24 in 10 seconds, with cues 3 = <4 in 10 seconds with cues 4 = cannot perform
Rigidity in arms	Right & left (n=2)	0 = absent 1 = slight or present only with activation 2 = mild to moderate 3 = severe, full range of motion 4 = severe with limited range
Bradykinesia in body	N/A (n=1)	0 = normal 1 = minimally slow (? normal) 2 = mildly but clearly slow 3 = moderately slow, some hesitation 4 = markedly slow, long delays in initiation
Gait	N/A (n=1)	0 = normal gait, narrow base 1 = wide base and/or slow 2 = wide base and walks with difficulty 3 = walks only with assistance 4 = cannot attempt
Tandem Walking	N/A (n=1)	0 = normal for 10 steps 1 = 1 to 3 deviations from straight line 2 = >3 deviations 3 = cannot complete 4 = cannot attempt

The UHDRS-TMS is widely used as an outcome measure and is currently considered the gold-standard for assessing motor symptoms in HD (Reilmann and Schubert, 2017a; Winder *et al.*, 2019). In a longitudinal study of the UHDRS assessments (motor, cognitive, behaviour & function) the UHDRS-TMS was found to be by far the most robust method for detecting change in early stage manifest HD (Meyer *et al.*, 2012). Similar results were found in a series of studies that showed over 12-, 24- and 36-month periods the UHDRS-TMS in manifest HD cohorts progressed significantly relative to controls and is related to functional decline over the same period (Tabrizi *et al.*, 2011, 2012, 2013). Notably however this effect was not been observed in pre-manifest participants and it has been suggested that other assessments that are sensitive motor symptoms and their progression within this cohort be developed (Tabrizi *et al.*, 2013).

The UHDRS-TMS has been criticised in the literature for ceiling effects in late-stage HD and a lack of sensitivity in early-stage HD where, although change can be detected, improved sensitivity is desirable (Youssov *et al.*, 2013; McColgan and Tabrizi, 2018). More specifically, the sensitivity of the 5-level rating used by the UHDRS motor items has been questioned along with the 'weighting' each item is given (e.g., chorea effectively has 7 times the weighting of bradykinesia which is only rated once) (Reilmann *et al.*, 2011a).

Whilst the UHDRS-TMS appears suitable to a degree for assessing motor symptoms in manifest HD, results throughout the literature suggest a more granular, sensitive method is required particularly for assessing pre-manifest, prodromal HD, and early-HD. This has led to the ongoing development of alternative methods of motor assessment, one such approach being instrumented assessments, the main focus of this thesis. Instrumented assessment of motor function along with other alternative motor assessments are discussed in section 1.4. Despite its flaws, the UHDRS-TMS utilises expert clinical knowledge and remains the gold-standard method of motor assessment for the HD population. As such, any new proposed assessment should be shown to have a strong relationship with the UHDRS-TMS or at least the UHDRS motor assessment item(s) the proposed assessment aims to assess or be sensitive to.

1.3.3.6 Assessing the functional domain: The UHDRS Function Capacity Assessment, Functional Assessment Scale, and Independence Scale

As mentioned in section 1.3.3.2, the UHDRS Functional Capacity Assessment, commonly referred to as the TFC, is the standard measurement used to assess function in HD (Wild and Tabrizi, 2014).

The TFC assess a person's impairment in five areas of daily living – occupation, finances, domestic chores, the level of care they require (e.g., at home care or nursing care) and general activities of daily living (ADL) (Kieburtz *et al.*, 1996). Each area is rated on either a 2- or 3-point ordinal scale with higher numbers indicating an increased level of autonomy and so greater functional capacity. The scores are then summed to produce a final total score, the TFC score which ranges from 0 to 13, with a higher score indicating greater functional capacity and so a lower disease stage. As discussed in section 1.3.3.2, the TFC is regularly used to sub-divide the manifest stage of HD into sub-stages, often called TFC stages (Wild and Tabrizi, 2014).

In addition to the TFC, the UHDRS also includes two additional functional assessments, the UHDRS Functional Assessment (FAS) and UHDRS Independence Scale (IS) (Kieburtz *et al.*, 1996).

The FAS, often thought of as a more detailed extension of the TFC, consists of 25 yes-no questions (to be answered by a clinician) which assess the patient's ability to perform various life tasks (e.g., "Could subject engage in any kind of volunteer or non-gainful work?") (Kieburtz *et al.*, 1996). The final FAS score is calculated by counting the number of 'yes' responses with higher scores indicating a greater level of autonomy.

The IS has clinicians rate a patient's independence based on 10 descriptors ranging from "No special care needed" to "Tube fed, total bed care". Each descriptor is assigned a score ranging from 100 to 10, clinicians can report scores ending in either 0 or 5 (e.g., 20 or 25) depending where on the scale they feel a patient fits (Kieburtz *et al.*, 1996).

The TFC, FAS and IS have all been used in numerous clinical studies and change over time has been reported for all measures (Mestre, Busse, *et al.*, 2018). However, the literature shows that these functional measures may nevertheless be insufficient for detecting progression during the earlier stages of HD.

The TFC was included in the battery of assessments used in the 12-, 24 and 36-month studies mentioned in the previous subsections (Tabrizi *et al.*, 2011, 2012, 2013). The authors concluded that whilst the TFC appears to be associated with whole and regional brain atrophy it is not sensitive enough for use in tracking progression in early manifest or pre-manifest populations (Tabrizi *et al.*, 2013). The TFC is also known to have a floor effect in early-stage HD, which is the period most treatments look to target (Mestre, Busse, *et al.*, 2018). It should however be recognised the despite its flaws the TFC is routinely used for the assessment of HD both in academic research and in clinical trials (Carlozzi *et al.*, 2014).

Similarly, a large-scale study of prodromal stage HD (i.e., HD-positive individuals close to motor manifestation) found that 88% of their cohort scored at the ceiling of both the TFC and FAS (Paulsen *et al.*, 2010). Whilst the FAS has been found to reliably degrade over time in early manifest HD, the mean annual rate of change is small (~0.95 points) (Meyer *et al.*, 2012). This rate of change equates to a just under single question being changed from 'no' to 'yes' on average per year which is unlikely to be sufficient for detecting small changes in response to pharmacological intervention.

The same study found that in early manifest populations the IS declined at a mean rate of approximately 3.21 points per year. This observed rate of decline along with the increased resolution of the IS relative to the TFC and FAS suggests it may make for a more sensitive outcome measure. However, whilst the IS is based on *expert* clinical opinion, it is still ultimately based on an opinion and so is inherently subjective. Additionally, there have been no studies that show the inter-rater reliability of the IS is sufficient for this subjectivity to be ignored (Mestre, Busse, *et al.*, 2018).

It is important to note that the TFC and FAS are also typically considered to be subjective as they are similarly opinion-based (albeit the opinion of expert clinicians) and no clinician-focused inter-rater reliability studies have been reported for either (Mestre, Busse, *et al.*, 2018).

As has been repeatedly stated, although the TFC, IS, and FAS are all opinion-based and therefore are by definition subjective, they are still the opinions of expert clinicians. This fact should not be discounted – the opinion of expert clinicians is invaluable and vital to the proper understanding an individual patient's disease state. However, the lack of study into inter-rater reliability coupled with their low sensitivity to early-stage HD is problematic when considering them as suitable for use as outcome measures (although again it should be noted that they often are used as outcome measures, particularly the TFC).

In general, we would argue that whilst these functional assessments are useful for monitoring high-level change over time and for clinical decision-making, they regardless have limited potential for detecting small, subtle changes. This has been evidenced by various previously cited studies that have in some manner explored the sensitivity of one or more of the TFC, FAS, and IS (Paulsen *et al.*, 2010; Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012; Mestre, Busse, *et al.*, 2018).

However, given the importance of assessing an individual's functional capacity for determining their quality of life it would be inadvisable to discount them entirely. As increased sensitivity has been reported for other measures one potential solution is to create composite measures by combining one or more of these functional assessments with assessments of other symptom domains. Such composite measures have recently received significant attention in the literature and are the topic of the next subsection.

1.3.3.7 Composite Measures

The assessments reviewed so far illustrate that changes in individual symptom domains as measured using gold-standard assessments can to one degree or another be observed over the course of HD. A clear limitation of tracking HD progression using singular symptom domains is however that it ignores the inherently multifaceted nature of the disease. Whilst one alternative may be to attempt to increase the sensitivity of measuring individual symptoms and so more finely track progression, another is to take multiple symptoms into account simultaneously by creating composite measures.

Currently there are two well-known composite measures developed for HD – the Composite Unified Huntington's Disease Rating Scale (CUHDRS) (Schobel *et al.*, 2017) and the Prognostic Index for HD

(PI_{HD}) (along with its normalised version the Prognostic Index for HD Normalised (PIN_{HD})) (Long *et al.*, 2017).

The CUHDRS is calculated by combining centred & scaled versions of the TFC, UHDRS-TMS, SDMT and SCWT, thus taking gold-standard measurements of function, motor, and cognition into account. The weights centring and scaling constants were calculated by Schobel *et al.*, (2017) based on a large sample (n=1668) of early-stage HD data. A higher CUHDRS indicates lower disease severity. Schobel *et al.*, (2017) found the CUHDRS declines at approximately 1 point per year in early manifest HD (TFC Stages 1 & 2) however the same was not seen in pre-manifest populations. It was found however that CUHDRS scores correlated with five disease groups – two pre-manifest groups (estimated to be either closer or further than 10.8 years from motor onset), TFC Stage 1, TFC Stage 2 and a healthy control cohort. More recently, Estevez-Fraga *et al.*, (2021) found that the CUHDRS was correlated with changes in brain volume across both pre-manifest and early-stage HD patients. Additionally, Estevez-Fraga *et al.*, (2021) found the correlation between the CUHDRS and changes in brain volume to be significantly stronger than the same correlations the TFC and TMS. In general, the CUHDRS has been shown to provide a more sensitive measure of clinical change in early manifest HD and has a superior relationship to structural brain changes than the other measures.

PI_{HD} combines the UHDRS-TMS, SDMT, Age and CAG to produce an index indicative of the risk of future motor diagnosis in pre-manifest HD and so provide an estimated rate of disease progression (Long *et al.*, 2017). As PI_{HD} values increase so too does the estimated risk of motor diagnosis. A normalised version of PI_{HD} named PIN_{HD} was developed to enhance interpretation by centring it around an estimated 0.5 probability of motor diagnosis within 10 years. Thus, a PIN_{HD} score < 0 indicates a greater than 50% chance of motor diagnosis within 10 years and a score > 0 indicates the opposite. In two of three external studies the results of Long *et al.*, (2017a) were confirmed, suggesting the use of PI_{HD} and PIN_{HD} in future studies looking at survivability, predicting progression and for more generally as inclusion criteria in longitudinal studies being worthwhile.

1.3.3.8 HD Assessment Summary

Assessing disease state in HD is important for two primary reasons. First, disease state assessment is crucial for proper clinical management, allowing informed decisions to be made regarding how medical professionals can best support individuals with HD, their families, and their carers (Clinch *et al.*, 2018). Secondly, in order for the efficacy of potential therapeutics to be shown during clinical trials, the symptoms caused by HD must be reliably detected and their severity & progression sensitively assessed & tracked (Reilmann and Schubert, 2017a).

Currently the gold-standard for assessing disease state in HD is accomplished using the UHDRS battery of assessments and the PBA-s. Based on the literature the suitability of these measures for tracking small changes over time, particularly in early HD, appears to be varied.

Whilst one set of studies found UHDRS measures of cognitive function to reliably change over time (Tabrizi *et al.*, 2011, 2012, 2013) a larger study of early manifest HD found the opposite, suggesting that at best they are not reliable for detecting change in early manifest HD (Meyer *et al.*, 2012).

Measures of functional capacity have been observed to reliably change over time in early manifest HD however it has been suggested the assessments used are too coarse for detecting subtle change over time (Tabrizi *et al.*, 2011, 2012, 2013). The use of functional measures in pre-manifest and prodromal HD appear to be limited with a dramatic ceiling effect having been observed in a prodromal cohort (Tabrizi *et al.*, 2013; Mestre, Busse, *et al.*, 2018).

Although the ability of cognitive and functional capacity measures to track changes over time seems to be limited, changes have still been found. Measures of behavioural abnormalities on the other hand have been universally found to be ineffective for detecting disease progression (Tabrizi *et al.*, 2011, 2012, 2013). This is the case to such an extent that they were excluded from the creation of the CUHDRS (Schobel *et al.*, 2017).

Motor symptoms as measured by the UHDRS-TMS were found to be reliable for measuring change over time in early manifest HD cohorts (Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012). Additionally, annual change was noticed in a prodromal cohort estimated to be close to motor diagnosis, however the scales suitability to measuring changes in pre-manifest participants appears to be limited (Tabrizi *et al.*, 2011, 2012, 2013). Additionally, concerns have been raised regarding its sensitivity, granularity, and the effective weighting it applies to different symptoms (Reilmann *et al.*, 2011a).

Given the multifaceted nature of HD, it is unsurprising that the composite measures seem to be superior for assessing HD disease state relative to their component assessments. The CUHDRS has been shown to outperform the individual UHDRS assessments for tracking changes in disease state and structural brain changes (Schobel *et al.*, 2017; Estevez-Fraga *et al.*, 2021) whilst PI_{HD} and PIN_{HD} both appear to be promising measures for estimating progression rates in pre-manifest and prodromal populations (Long *et al.*, 2017). Composite measures also have the advantage of giving a more complete ‘snapshot’ of HD disease state than any measure of singular disease symptoms, taking into account multiple symptom domains simultaneously.

Unlike the traditional gold-standard measures however the CUHDRS, PI_{HD}, and PIN_{HD} have not been robustly explored in numerous longitudinal studies or in large samples of specific cohorts although such work is being conducted. Additionally, a limitation of both these measures is that they rely on existing assessments that are known themselves to lack sensitivity. For example, whilst the CUHDRS may be a better global indicator of disease progression than the UHDRS-TMS, if a patient's motor symptoms progress to a degree undetectable by the UHDRS-TMS then the CUHDRS will register no change either. In short – whilst composite measures may be more sensitive than their component measures, they will still be limited by the sensitivity of those component measures.

Overall, the literature points towards composite measures and motor symptom assessment being the most immediately promising avenues for continued development. Large-scale longitudinal studies will be required to further develop the CUDHRS, PI_{HD} and PIN_{HD} which is out of the scope of this thesis.

The continued development of motor symptom detection and estimation however is ripe for investigation. HD is recognised as a motor disorder

The continued development of motor symptom detection and estimation however is ripe for investigation. HD is recognised as a motor disorder, with motor symptoms being among the first symptoms to present (Reilmann and Schubert, 2017a). As such, the UHDRS-TMS is used as an outcome measure for clinical trials investigating both generic and motor-specific HD therapies (Reilmann and Schubert, 2017a; Winder *et al.*, 2019). The UHDRS-TMS is however known to be insensitive during early-stage HD (Youssov *et al.*, 2013; McColgan and Tabrizi, 2018). Parallel to this, as has been discussed previously (see section 1.3.1), one particularly promising avenue of research seeks to inhibit the production of the mHTT protein (Wild and Tabrizi, 2017). Due to the progressive damage caused by the mutant protein, such therapies will be most effective if applied during the pre-manifest phase of the disease before overt symptoms present (Rosser and Svendsen, 2014). Thus, by developing assessments sensitive to the progression of early-stage motor symptoms, such as chorea, the evaluation and ultimately the development of potential therapeutics can be aided.

The next section of this chapter focuses on alternative methods which have been suggested for assessing motor symptoms in HD. Particular attention will be paid to methods which have made use of modern sensor technology to directly measure such symptoms, commonly referred to as *instrumented clinical assessments* or simply *instrumented assessments*.

1.4 Part 2: Alternative measures of motor function

1.4.1 Alternative measures of motor function: Overview

The sensitive assessment of motor symptoms in HD is crucial to proper clinical management and ongoing clinical trials. There is however an ongoing issue regarding the sensitivity of the UHDRS-TMS, the gold-standard assessment used to assess HD movement disorders. Improving the sensitivity of HD motor assessments will not only aid clinical management and clinical trials but also composite measures of disease state, such as the CUHDRS, which incorporate the UHDRS-TMS. As a result, numerous alternatives have been suggested, ranging from modifications to the UHDRS-TMS (Youssov *et al.*, 2013) to novel instrumented assessments (Reilmann *et al.*, 2011a).

The following sections split alternative motor assessments into one of two types – traditional clinical assessments and instrumented clinical assessments. The distinction is that instrumented clinical assessments use modern sensor technology (e.g., accelerometers, inertial measurement units, etc) to collect additional data whereas traditional clinical assessments typically rely on simpler scoring methods (e.g., timing of tasks, number of correct answers, observer ratings, etc). For the sake of brevity, *traditional clinical assessments* and *instrumented clinical assessments* are usually referred to throughout this thesis as *clinical assessments* and *instrumented assessments* respectively.

Due to the underpinning rationale of updating motor assessments in HD to aid clinical management and clinical trials, particular attention is paid to discussing the assessments suitability for reliably detecting change over time in early-stage HD cohorts.

1.4.2 Traditional clinical assessments of motor function

There have been many alternative clinical assessments of motor function proposed for HD. Due to the breadth of work in this area, the International Parkinson and Movement Disorder Society commissioned a review into the various clinical rating scales (i.e., those reliant on expert clinical opinion) of motor function that have been used in HD (Mestre, Forjaz, *et al.*, 2018). The review initially identified 27 such scales with 6 of them ultimately being retained after exclusion criteria were applied. It is notable that a large portion (n=16) of the identified scales were removed as they were entirely based on subsets of the UHDRS motor section (see supplementary of Mestre, Forjaz, *et al.*, (2018)). Each of the retained items were then assessed using the criteria below (adapted from Mestre, Forjaz, *et al.*, (2018)).

1. Scale has been used in HD populations.

2. Scale has been used in HD by groups other than its original developers and data of its use were available. If the scale was not originally developed for HD, then this criterion was met if at least one group had used it with HD and reported clinometric or psychometric data on HD using it.
3. The available clinimetric or psychometric data in HD support the goals of screening or measurement of severity of motor function.

Based on the above criteria, a scale recommendation level of recommended, suggested, or listed, was assigned as shown in Table 7. Of the 6 retained scales 5 were listed as recommended or suggested. The recommendation level these 5 scales is shown in Table 8.

Table 7: Recommendation levels & criteria from Mestre, Forjaz, *et al.*, (2018)

Recommendation	***Criteria Requirements***
Recommended	*(1) and (2) and (3) met*
Suggested	*(1) and either (2) or (3) met*
Listed	*(1) met only*

Table 8: Rated alternative motor assessments from Mestre, Forjaz, *et al.*, (2018)

Scale	***Recommendation Level***
UHDRS-TMS	*Recommended for assessing severity of motor symptoms*
UHDRS-TMS4	*Suggested for assessing severity of motor symptoms*
Quantified neurological Examination	*Suggested for assessing severity of motor symptoms*
Marsden and Quinn Chorea Severity Score	*Suggested for assessing severity of* ***chorea*** *symptoms*
Abnormal Involuntary Movement Scale	*Suggested for assessing severity of* ***chorea and dystonia*** *symptoms*

Of the retained scales, only the UHDRS-TMS met the criteria for being recommended as an assessment of motor symptom severity and none met the criteria for recommendation for screening or assessing change over time. Based on the findings of the review, the authors conclude that out of the available assessments of motor symptoms in HD the UHDRS-TMS is the one best suited for use in clinical practice and research purposes. They also conclude, however, that there is a clear need for tools to be developed for detecting and assessing subtle manifestations of motor symptoms in HD, which they felt none of the reviewed assessments were suitable for. In terms of alternative clinical assessments of motor function currently available in HD, Mestre, Forjaz, *et al.*, (2018) succinctly illustrates that whilst many may have been developed none currently surpass the UHDRS-TMS.

1.4.3 Instrumented clinical assessments of motor function

As was demonstrated in the previous section, the UHDRS-TMS is widely accepted as the tool best suited for assessing motor function in HD of those currently available. However, as has also been also demonstrated, the UHDRS-TMS suffers from two widely recognised drawbacks – a lack of sensitivity to change over time and ill-suited for detecting subtle symptoms. One alternative, as stated by Mestre, Forjaz, *et al.*, (2018), is to incorporate modern-day sensor technology into existing & novel clinical assessments, to create instrumented clinical assessments (a.k.a. instrumented assessments).

Instrumented assessments have been applied in numerous domains using a multitude of different technologies for many years, especially for assessing motor function/symptoms. However, only a limited number of such assessments have been developed for HD. As such it is useful here to also consider instrumented assessments in PD where there is a much wider body of work relative to that in HD, likely due to PD's significantly higher prevalence in the population (PD: 1-2 per 1000; HD: 6-13 per 100,000) (Rawlins *et al.*, 2016; Tysnes and Storstein, 2017).

The rationale for developing instrumented assessments of motor function in PD is similar to that in HD – the gold-standard clinician-rated motor assessment scale in PD (the Unified Parkinson's Disease Rating Scale (UPDRS)) is thought to be coarse with limited sensitivity to change over time (Clarke, 2007). Although not all motor symptoms are shared between HD and PD there are some common symptoms including rigidity, bradykinesia and, more generally, hypokinesia (i.e., poverty of movement). Thus, as the rationales are equivalent and the movement disorders similar, the techniques used to develop instrumented assessment in PD will likely be translatable for developing similar assessments in HD.

The remainder of this section is split into two parts. First, there is a broad overview of instrumented clinical assessments in PD using a recently conducted wide-scale review. The aim of this first section is to understand the typical trends and recommendations for instrumented assessment development that have come out of PD research. Secondly, the trends and notable instrumented clinical assessments that have been developed for HD are stated discussed.

1.4.3.1 Instrumented clinical assessments: Parkinson's Disease

The number of instrumented assessments developed for PD is vast. A 2017 review found 1429 articles written between 2006 and 2016 (of which 136 were retained after exclusion criteria were applied) that looked at instrumented assessment for a number of applications in PD (Rovini, Maremmani and Cavallo, 2017). The review broke down the literature into five applications – early diagnosis, tremor detection & severity estimation, body motion analysis, motor fluctuations & on/off phases, and home/long-term monitoring. Whilst all these applications do not directly relate to the goal of detecting and estimating motor symptom severity, they are nonetheless informative as to how instrumented assessments may be developed effectively to produce useful measures.

A variety of sensor types were used across the different applications and studies however accelerometers were by far the most common (64 papers), followed by gyroscopes (33 papers), and then EMGs (12 papers). The reviewed studies typically followed the process of having participants perform some type of action task whilst wearing one or more sensors. Once data collection was completed, features were extracted, and various methods used to relate them to gold-standard measures. As an aside, '*features'* is a term widely used in machine learning to denote individual, measurable quantities that in some way describe a phenomenon being studied. Feature extraction is the process of extracting features from said phenomenon and may be sometimes referred to as feature engineering. A simple example of feature extraction might be calculating the mean, minimum and maximum acceleration (the features) from a recorded acceleration signal (the phenomenon). These features can then be used in various types of analyses, ranging from simple correlations and descriptive statistics to variables in regression analysis and machine learning classifiers, to inputs for neural networks.

A large variety of features were used throughout the studies although the use of frequency-domain features (e.g., dominant/median frequency, spectral edge estimates) and time-domain features (e.g., root-mean-square of acceleration, peak & average velocity, various measures of jerk) were particularly common. The method used to associate features to gold-standard measures varied depending on the application. Whilst some advanced techniques were used (e.g., Support Vector Machines (SVMs), neural networks), more traditional statistics were far more common (e.g., ANOVAs, Mann-Whitney U, correlation statistics).

From the trends identified by Rovini, Maremmani and Cavallo, (2017) and the associated general discussion, there are a number of takeaway points that can be applied in HD. First, body-worn inertial sensors are suitable for assessing motor symptoms. Second, extracted sensor features should be tailored to the individual application. Third, any proposed instrumented assessment should be related back to existing gold-standards as evidence of validity. These will now each be discussed in turn in more detail.

Based on the reviewed literature, one approach for measuring motor symptoms would be to use a set of portable, cost effective, wearable inertial sensors with a high sampling rate (one suggestion being 100Hz). Such a tool would be suitable for use in both small-scale and large-scale studies whilst also being cost-effective for eventual use in health services and clinics. These use-cases just as relevant in HD as they are in PD and so should be taken into consideration when selecting sensor types.

Regarding the recommendation of inertial sensors, whilst they would not be suitable for assessing all motor symptoms in HD (e.g., oculomotor dysfunction) they would likely be well suited for assessing motor symptoms such as chorea, rigidity, and bradykinesia. Whilst more accurate sensors (specifically full-body motion capture) are available, the price-point, setup time and high-level of expertise required to correctly (and consistently) operate such technologies relative to the accuracy increase they may provide makes their use difficult to justify. An additional difficulty in HD, and to an extent PD, with such marker-based technologies is involuntary movements seen in both diseases. A calibration phase is usually required for marker-based motion capture which requires participants to remain motionless – this will obviously be problematic in symptomatic HD and PD and may significantly impact data quality.

It is additionally suggested that results be computed either on the device itself or transmitted to a nearby control station. Such considerations should not impact the initial design of an instrumented assessment where the primary concern should be the validation and fine-tuning the assessment itself (e.g., construct validity and sensor set reduction). The requisite components for processing results either onboard or remotely would be developed naturally over the course of the assessments design and validation phases with only minor alterations being required to optimise the delivery of results if such efficiency were required.

Once data has been collected features tailored to the individual application should be developed and extracted. Whilst there were some clear trends in the features used across the different applications which ones should be used remains open for debate. Whilst we cannot simply take a small set of the features discussed by Rovini, Maremmani and Cavallo, (2017) and see if they are also applicable in HD, this does highlight two important points about featuring engineering - it is more of an art than it is a science, and it is vital when developing instrumented assessments. Whilst conceptually simple (think of features that can be extracted, extract them, and then observe their relationship with whatever you are trying to measure/model), the difficulty lies in knowing which features should be extracted. Essentially, feature engineering is the translation of (ideally expert) domain knowledge into discrete quantifiable measurements that can be used to study a phenomenon using whatever analytical techniques are appropriate for answering the question(s) at hand. This can however lead to the assumption that when we do not know which features to extract for a given application, the most appropriate course of action is to generate as many features as possible.

With modern computing, a very large number of features can be relatively quickly generated and tested (either in isolation or at the extreme end even in a single model). This is well illustrated when one considers a device such as a triaxial accelerometer. A single one of these devices will produce three signals (acceleration along the x-, y- and z-axis) which can be made into an additional four signals (xy vector, xz vector, yz vector and xyz vector (a.k.a. absolute acceleration)). Thus, if a feature is extracted per-signal, a single recording of a single sensor could potentially generate up to 7 instances of every type of feature a researcher wishes to extract.

Although this proposition initially seems sensible and may even appear very efficient it is inherently flawed. Having a very large feature space can not only drastically increase computation time but more importantly may easily lead to the overfitting of models and increase the probability that detected relationships (e.g., correlations) are actually false positives (i.e., a type I error). Methods like cross-validation, feature selection, and post-hoc adjustments (e.g., Holm-Bonferroni corrections) may be used to minimise these concerns. Whilst such methods should always be applied when multiple features are generated there is a clear advantage in being as selective as possible about which features are extracted from the outset. In the context of developing instrumented assessments for HD, there is clearly a rationale for closely analysing the clinical description of motor symptoms seen in HD, as well as utilising expert clinical knowledge when deciding which features to extract.

Additionally, as is also pointed out by Rovini, Maremmani and Cavallo (2017), any features that are generated must be as explainable as is practical. To an extent this is just as much a consideration when thinking about which features to extract as the theoretical suitability of the features themselves. Acceptance and uptake of a proposed assessment will likely be hindered if the features it is based around are not readily explainable to clinicians, medical practitioners, the wider research community, and ideally the patients themselves.

Finally, if an instrumented assessment is to be accepted a logical course of action is to relate it to the existing gold-standard(s). The vast majority of studies reviewed have attempted to relate measures to the MDS-UPDRS III (effectively the PD parallel of the UHDRS-TMS). A significant relationship between an instrumented assessment and the MDS-UPDRS III provides additional confidence and evidence that the features and assessment in use are sensitive to the underlying motor symptom(s). The same technique of providing evidence of instrumented assessment validity can of course be applied in HD, using either the UHDRS-TMS as a whole or specific motor items (e.g., the chorea, bradykinesia, or dystonia UHDRS-TMS sub-items). A natural extension of this idea is the validation of features directly by one (or ideally) more clinician (if possible). For example, if an algorithm purports to detect individual choreatic movements, then recording the patients during the assessment and having expert clinicians also note which movements they thought were choreatic, and then comparing the two, would give insight into the algorithm's validity.

In general, the wide-ranging body of research that has been conducted on instrumented assessments for PD is highly informative about the general approach that may be taken when one looks to develop similar assessments for HD. This is particularly the case given that the limitations of the MDS-UPDRS III are generally the same as those found in the UHDRS-TMS – subjective, ordinally rated scales which may lack in sensitivity, granularity, and the clear potential for interrater variability (Reilmann *et al.*, 2011a; Rovini, Maremmani and Cavallo, 2017). Again, it should be noted that these critiques do not in any way invalidate the existing gold-standards and instrumented assessments should not be viewed as an attempt to replace the expert clinical opinion these gold-standard are based on. The benefit that instrumented assessments can provide to clinicians is an enhanced picture of the symptom's patients display by utilising modern sensor technology. This relationship is similar to that between traditional statistics and the rise of 'machine learning' techniques that are currently highly prevalent throughout general academic literature. Machine learning complements traditional statistics by providing additional evidence, viewpoints, and functionality but it does not replace them.

1.4.3.2 Instrumented assessments: Huntington's Disease

The same breath of work for instrumented assessment in PD does not exist in HD, although there still are some relevant examples (Bechtel *et al.*, 2010; Reilmann, Bohlen, Klopstock, Bender, Weindl, Saemann, Auer, E. Bernd Ringelstein, *et al.*, 2010; Reilmann, Bohlen, Klopstock, Bender, Weindl, Saemann, Auer, Erich B. Ringelstein, *et al.*, 2010; Reilmann *et al.*, 2011a; Mann *et al.*, 2012; Dalton *et al.*, 2013; Mannini *et al.*, 2015, 2016; Gwin *et al.*, 2016; Reilmann and Schubert, 2017b; Kegelmeyer *et al.*, 2017; Acosta-Escalante *et al.*, 2018; Bennasar *et al.*, 2018; Jensen *et al.*, 2018; Purcell *et al.*, 2019; Dinesh *et al.*, 2019; Gordon *et al.*, 2019; Gaßner *et al.*, 2020). As this thesis focuses on HD, and as comparatively to PD there are few relevant studies, these studies will be discussed more specifically than in the previous section. Due to the importance the C3t (and its instrumented variant) in this thesis, this section is further divided into two subsections – the first on general trends of instrumented assessments for HD and the second on the instrumented C3t.

1.4.3.2.1 General Trends

As in PD, the reviewed HD studies make heavy use of accelerometers and IMUs with many of the developed instrumented assessments making use of one or more such sensors (Reilmann *et al.*, 2011a;

Mann *et al.*, 2012; Dalton *et al.*, 2013; Mannini *et al.*, 2016, 2015; Kegelmeyer *et al.*, 2017; Acosta-Escalante *et al.*, 2018; Bennasar *et al.*, 2018; Jensen *et al.*, 2018; Purcell *et al.*, 2019; Dinesh *et al.*, 2019; Gordon *et al.*, 2019; Gaßner *et al.*, 2020). A longitudinal study found people with HD, including those with moderate cognitive symptoms, considered accelerometers to be easy to use and comfortable enough to continuously wear them over long periods of time (Andrzejewski *et al.*, 2016).

Also similar to PD, the developed assessments tended to focus on a combination of showing statistical differences found between assessment measures for patient groups and/or showing statistical relationships between said measures and accepted gold-standards. Unlike in PD however there appears to be less of a focus on individual symptoms in favour of assessing general motor dysfunction, usually via relationships with the UHDRS-TMS. The exception to this is the assessment of gait and postural stability which a number of papers are based around (Dalton *et al.*, 2013; Mannini *et al.*, 2015, 2016; Kegelmeyer *et al.*, 2017; Acosta-Escalante *et al.*, 2018; Jensen *et al.*, 2018; Purcell *et al.*, 2019; Gaßner *et al.*, 2020), and several of the Q-Motor assessments (Reilmann and Schubert, 2017b).

Whilst many studies, especially those exploring instrumented assessment of gait, utilise simple, easy to explain features, several make use of much more complicated techniques (Mann *et al.*, 2012; Mannini *et al.*, 2016; Acosta-Escalante *et al.*, 2018; Bennasar *et al.*, 2018; Gordon *et al.*, 2019). Although effective, the added complexity and resulting difficulty in providing clear clinical translation may limit uptake as discussed in section 1.4.3.1.

Data analysis is typically limited to standard statistical techniques (e.g., correlations, group differences, effect sizes) although there is some use of more advanced machine learning techniques. A particular lack of machine learning techniques applied in HD, unlike in PD, was noted by Bennasar *et al.*, (2018a). This is unsurprising however considering the reduced amount of work on instrumented assessment conducted in HD overall, and the widespread use of machine learning techniques being a relatively trend.

Out of the various instrumented assessments developed for HD the Q-Motor series of assessments is arguably the most well-known and widely applied. Q-Motor, which stands for Quantitative Motor, was originally developed, as were many of the other assessments, in response to the insensitivity of the UHDRS-TMS. Q-Motor consists of 6 standard tests - digitomotography, manumotography, choreomotography, glossomotography, dysdiadochomotography, and pedomotography (Reilmann and Schubert, 2017b).

Digitomotography is a speeded tapping test where participants use the index finger of their non-dominant hands to tap a force sensor (Bechtel *et al.*, 2010). Two tasks are conducted; a speeded tapping task where participants tap the sensor as fast as possible for a set duration and a metronome period task where participants tap in time to an auditory metronome. Numerous features are extracted from the resultant signals, including the variability of tap durations, the interonset-, interpeak- and intertap-intervals, the variability of peak tapping forces, and tapping frequency.

Manumotography has subjects grasp, lift, and attempt to hold stationary a weighted object equipped with two force-torque sensors (Reilmann, Bohlen, Klopstock, Bender, Weindl, Saemann, Auer, Erich B. Ringelstein, *et al.*, 2010). Extracted features include grip forces, grip force variability, the time between contact of thumb and index finger, duration between contact time and positive load force initiation, time between load force onset to pick-up, and time between onset of lift to maximum grip force.

The weighted object used for manumotography is also equipped with an inertial sensor suitable for assessing position and orientation of the object allowing it to be used for the choreomotography test. The choreomotography test captures position and orientation data of the weighted object whilst it is held stationary for 35 seconds. Extracted features include changes in the position (derivatives of x-, y-, z- axes calculated to produce velocity, the mean absolute value of each is taken and then all values are summed) and changes in orientation (mean of absolute values of the derivatives of roll, pitch, and yaw are calculated and then summed) (Reilmann *et al.*, 2011a).

Glossomotography assesses force coordination during tongue protrusion, i.e., the ability of a participant to apply a consistent force using their tongue (Reilmann, Bohlen, Klopstock, Bender, Weindl, Saemann, Auer, E. Bernd Ringelstein, *et al.*, 2010). Participants are seated in front of a force transducer and a monitor is used to display the current force being applied to the transducer as well as a line indicating the desired force level. The test has participants use their tongue to apply enough force to the transducer such that the desired force level is reached and maintained. The test is repeated three times with different target force levels (0.25 N, 0.5 N, and 1.0 N). Extracted features include mean & variability of the tongue protrusion forces, percentage of time tongue protrusion forces remained at desired level, and tongue contact time. The final two tests, dysdiadochomotography (assesses regularity of alternating pronation/supination hand tapping) and pedomotography (assesses regularity of foot tapping) do not appear to be covered in published works and are only referenced when Q-Motor is discussed more generally (Reilmann and Schubert, 2017b).

The early Q-Motor papers have shown the various assessments to be linked to gold-standard clinical measures. Studies have shown significant Pearson's R correlations between the UHDRS-TMS and features derived from digitomotography (r=0.67), choreomotography (r=0.73), manumotography (r=0.61), and glossomotography (r=0.68) in both manifest and pre-manifest HD cohorts (Bechtel *et al.*, 2010; Reilmann, Bohlen, Klopstock, Bender, Weindl, Saemann, Auer, E. Bernd Ringelstein, *et al.*, 2010;

Reilmann, Bohlen, Klopstock, Bender, Weindl, Saemann, Auer, Erich B. Ringelstein, *et al.*, 2010; Reilmann *et al.*, 2011a). Similar results have been reported for the disease burden score (DBS), a measure of general disease state based on the CAG repeat count and age (Penney *et al.*, 1997; Reilmann and Schubert, 2017b). Additionally, digitomotography has been found to be related to the UHDRS measure of finger tapping (r=0.65) and choreomotography has been found to be linked, albeit quite weakly, to a UHDRS measurement of whole-body chorea (r=0.46). It should be noted however that whether latter, choreomotography, is actually measuring chorea has been debated in a letter to the editor of Movement Disorders (Casula *et al.*, 2018). The argument was that the features measured during choreomotography will be affected by multiple types of involuntary movement not just chorea, and that it is more a measure of how well an adopted posture can be maintained than a measurement of chorea.

The TRACK-HD series of studies found the Q-Motor assessments to be linked to progression in manifest HD cohorts over 12-, 24-, and 36-months (Tabrizi *et al.*, 2011, 2012, 2013). Digitomotography appeared to be particularly sensitive to progression, being one of only three measures to show progression in pre-manifest populations over 24-months and the only non-imaging measure to show changes in very early pre-manifest HD (>10.8 years from estimated motor onset diagnosis) over a 36-month period (Tabrizi *et al.*, 2012, 2013).

Overall, the Q-Motor assessments have been shown to be useful for objectively estimating motor function, are related to the degeneration of striatal volume, and reliably track changes over time in manifest HD cohorts. However, although the Q-Motor assessments have been repeatedly shown to have a strong relationship with the UHDRS-TMS there has been comparatively little work in establishing their relationship with individual motor symptoms. The exceptions being digitomotography (Pearson's R = 0.65 with UHDRS-TIMS finger tapping sub-item) and choreomotography (Pearson's r = 0.46 with UHDRS-TMS whole-body chorea sub-item).

Although the Q-Motor assessments are clearly linked with motor function in HD and illustrate the benefits instrumented assessments can provide, they clearly do not solve the problem of sensitively measuring/assessing specific HD motor symptoms.

1.4.3.2.2 The Instrumented Clinch Token Transfer Test

A more recent attempt at objectively measuring motor symptoms in HD, and one which is inexorably linked to this thesis, is the work conducted by Bennasar *et al.*, (2018a) who instrumented the Clinch Token Transfer Test (C3t) with accelerometers with the aim of objectively measuring upper-limb motor function in HD.

As mentioned in section 1.2, the C3t is a multi-stage token transfer assessment which has participants transfer a series of tokens from their initial position on a board into a slotted box.

Bennasar *et al.,* instrumented the C3t by having participants wear tri-axis accelerometers whilst completing the tasks – 2 on their wrists (1 on each wrist) and 1 strapped to their sternum. The resultant acceleration signals (9 in total) were then extracted, and 234 features generated using the signals. These are listed in Table 9, sub-divided into frequency domain features and time domain features. Feature selection was employed to determine the most relevant features for distinguishing between healthy controls and manifest HD participants. The five most relevant features were then used as dependent variables for a linear regression model with a summary score of upper limb function, the Modified Upper-limb Motor Score (MULMS), used as the dependent variable. The MULMS was generated by summing five scores from the UHDRS-TMS (left & right upper-limb dystonia, left & right upper-limb chorea, and trunk chorea).

Table 9: Features generated from acceleration signals by Bennasar *et al.*, (2018a)

Frequency Domain Features
Feature
Average magnitude of 5 SFFT
Frequency Domain Entropy
Magnitude coefficients of wavelet
Spectral Energy
Wavelet Energy
Wavelet Entropy
Time Domain Features
Feature
Average diagonal line
Determinism
Lyapunov Exponent
Mean Correlation between axis
Permutation Entropy

Recurrence Entropy
Recurrence Rate
Standard Deviation

The results reported by Bennasar *et al.*, were statistically significant; the Pearson's R correlation between the value generated by the regression model and the MULMS was 0.77 (r^2=0.59, p<0.01) and the Mean Absolute Error (MAE) was 2.1 points (Normalised MAE of 12.4%). Whilst the results were positive however the work it does suffer from the issue of numerous highly complex features, the downsides of which have been mentioned throughout section 1.4.3. Additionally, as with many of the Q-Motor assessment, the developed measure relates to a composite of upper-limb motor function formed of chorea and dystonia rather than single motor symptom. Overall, Bennasar *et al.*, shows that the C3t may have potential for assessing upper-body motor symptoms, but further work is required.

1.4.4 Alternative measures of motor function: Summary

Effective clinical management of HD relies on correctly assessing symptom severity. Whether the goal is noticing the gradual emergence of symptoms during the pre-manifest phase, understanding the dominant symptoms present during the early manifest stage, or tracking the progression and evolution of symptoms as the disease progresses. Doing so allows clinicians to perform a wide variety of tasks necessary for making appropriate care recommendations. Similarly, when developing interventions, the progression of symptoms (or ideally lack thereof) is often used as evidence of an intervention's effectiveness. As the sensitivity of methods used to assess symptoms goes up, so to the minimal detectable effect of an intervention on those symptoms go down.

The UHDRS-TMS is the current gold-standard assessment of motor symptom severity in HD and is reliant on observation using the human eye. Such observation is based around expert clinical opinion but is there is ultimately a limit to the subtlety of movement that can be detected in this manner, one that is likely far lower than that of sensor technology. Experimental evidence has shown that whilst the UHDRS-TMS is suitable for detecting change its sensitivity is insufficient for pre-manifest & prodromal participants and floor effects occur in early-stage HD. Additionally, a recent review of the literature has shown that no viable alternative to the UHDRS-TMS currently exists.

In PD there are similar limitations with its gold-standard assessment of motor function, the MDS-UPDRS III. As a result, numerous researchers have developed instrumented assessment of movement disorders many of which have been highly successful. Although similar work has been conducted in HD many of the developed instrumented assessments focus on estimating overall motor symptom severity via the UHDRS-TMS score rather than specific motor symptoms. The core problem with this strategy is that estimating the UHDRS-TMS as a whole does not solve the need for more sensitive assessments of early-stage symptoms. The assessment of early-stage symptoms is critical to the continued development of potential disease modifying therapies for HD, many of which aim to target the earliest stages of the disease.

There is a clear gap in the literature for a sensitive instrumented assessment for specific, early-stage HD motor symptoms. As such an assessment would be targeted to specific symptoms, it can potentially be developed to be easier to explain than a more general one, as it is likely to be simpler to explain why a given assessment assesses an individual symptom type than why one assesses all symptoms covered by the UHDRS-TMS combined. This quality is desirable as it may work to enhance clinical uptake as discussed in section 1.4.3.1. In HD instrumented assessments of specific symptoms have so far been primarily developed for gait and postural stability as discussed in section 1.4.3.2.1. The exception to this is the Q-Motor battery where several assessments have shown links to specific motor symptoms. The choreomotography assessment however, although found to be related to chorea, is at best only moderately so (Pearson's R = 0.46), and the validity of the findings has been publicly questioned, again as discussed in 1.4.3.2.1.

In summary, based on the literature reviewed it can be concluded that there is currently an unmet need for assessing specific motor symptoms in HD using an instrumented assessment. Section 1.5 will utilise the clinical understanding of HD presented in section 1.3 and the lessons learned from similar studies in section 1.4 to discuss the design considerations of an instrumented assessment for motor symptoms in HD.

1.5 Part 3: Designing an instrumented assessment for motor symptoms in Huntington's Disease

1.5.1 General overview

As was shown throughout section 1.4 there is a need for instrumented assessments of motor function in HD based on wearable sensor technology, which is the core topic of this thesis. Parallel to this as mentioned in section 1.2, during the course this project the opportunity arose to vastly increase data collection by embedding the developed instrumented assessment into two additional studies. Doing so required the development of a remote data collection platform (RDCP) to facilitate data collection.

RDCPs and wearable technology are increasingly important in medical research. A recently published study (Bakker *et al.*, 2019) found 275 individual publications that included the development of mobile technologies and associated software systems for use in clinical research. This area has become of such great import that the Clinical Trials Transformation Initiative has been set up with the purpose of publishing and disseminating formal recommendations for the development of such technology (Initiative, 2018).

Considering the importance such systems have in medical research, and the fact that the development of the RDCP ultimately formed a substantial part of this project, the design considerations for the RDCP as well as the wearable technology used for the developed instrumented assessment are covered in this section. Covered areas include the design considerations of the instrumented assessment and RDCP artefacts, the various high-level decisions made regarding them during this project, and relevant background information on applicable techniques & technologies.

1.5.2 Instrumented Assessment Design

1.5.2.1 Instrumented Assessment Design: Overview

The design of an instrumented assessment is broken down here into 3 core parts. First, the requirements of an instrumented assessment with respect to assessing motor symptoms in HD are discussed. Secondly, the components which make up an instrumented assessment are detailed along with the various design choices made throughout this project. Finally, the analyses required to show an instrumented assessment's efficacy and the methods that may be used to do so are introduced.

1.5.2.2 Instrumented Assessment Design: Requirements

The requirements of an instrumented assessment can be split into two types, general requirements applicable to any instrumented assessment and specific requirements for the specific instrumented assessment use-case. These requirements drive the specific design considerations of any instrumented assessment and those outlined here will be used to guide the decisions made during section 1.5.2.1.

General design requirements require no specialised knowledge to uncover. A well-designed instrumented assessment should be cost effective, require as little training as possible to use, and be resistant to user error. These qualities are particularly important for instrumented assessments intended for medical research given that ideally the developed assessment will eventually be rolled out in large-scale studies.

The more specific requirements of an instrumented assessment will vary by use case. The use-case of this thesis is to assess/estimate motor function in HD and so the instrumented assessment designed such that it is suitable for assessing these types of symptoms. Additionally, the sensor features extracted from the developed assessment will ultimately be linked to motor symptoms and so should ideally be natural to relate to the current clinical understanding of HD motor symptoms. The reason for this, as discussed during section 1.4.3, is to enhance clinical uptake and to provide additional assurance of construct validity.

As was also discussed in section 1.4.3 whilst complex feature extraction techniques can be highly effective, they are not necessarily the best option. The main downside is that highly complex features can make the developed assessment a so called 'black box' – a piece of technology that works, but no one knows why. Black box technology can be highly effective, however in medicine the understanding of why & how technology works is crucial given what is at stake if the technology is flawed. Similarly, black box technology cannot provide evidence of construct validity – if it is not understood how the technology underpinning an instrumented clinical assessment works then it cannot be explained why that assessment is assessing the phenomena it is designed to assess.

Simply put, if the goal of an instrumented assessment is ultimately to provide evidence of success or failure during a clinical trial, the results are much easier to trust and the evidence all the more convincing if the instrumented assessment used to assess it is understandable.

Overall, with respect to this project the developed instrumented assessment was designed to meet 5 requirements.

1. The instrumented assessment must be shown to be sensitive to motor function in HD
2. Simple feature extraction techniques easily linked to symptoms should be preferred
3. User errors should, where possible, be mitigated by design choices
4. The instrumented assessment should be cost effective, or capable of becoming so
5. Minimal training should be required to correctly operate the instrumented assessment

1.5.2.3 Instrumented Assessment Design: Components

1.5.2.3.1 Overview

Instrumented assessments can be split into 3 main components – the underlying task/assessment (e.g., holding a weight in choreomotography, transferring tokens in the C3t), the sensors used to capture data during performance of the task/assessment and the features extracted from that data that are to be linked to disease state. The various design considerations that were considered and the decisions made with respect to this project for each of these components are discussed in the following sections.

1.5.2.3.2 Underlying Task/Assessment

1.5.2.3.2.1 General Considerations

When developing an instrumented assessment, a choice can be made as to whether an existing traditional clinical assessment should be instrumented, or an entirely new assessment created. For example, choreomotography is an entirely new instrumented assessment whereas the C3t is an existing clinical assessment which can be instrumented.

The main advantage of developing an entirely new assessment is that every aspect of the assessment can be tailored to the assessments use-case, potentially increasing the sensitivity of the assessment. The main disadvantage however is that it will lack an established body of literature to build upon. Previous literature can be particularly useful in cases where the instrumented version will serve similar use cases to the non-instrumented version, as that literature can act as a benchmark for the success of the instrumentation.

In the original project specification, which resulted in this thesis, the decision was made to focus on instrumenting an existing clinical assessment, the C3t. The C3t was felt to be a good candidate for instrumentation as it has been previously shown to be related to a variety of gold-standard HD clinical assessments including the SDMT, SWRT, UHDRS-TMS and TFC (Clinch *et al.*, 2018). Additionally, the test protocol (covered at length in the next section) lends itself naturally to being instrumented. Notably whilst the C3t's relationship with whole-body motor function was explored by Clinch *et al.*, (2018) using the UHDRS-TMS, its relationship with individual motor symptoms such as chorea was not. Additionally, as mentioned previously (Bennasar *et al.*, 2018) instrumented the C3t to assess generic upper-body motor function which, whilst a different use case to the one presented here, suggests that instrumenting the C3t is viable and worth exploring further. Finally, the C3t was originally designed to be a measure of fine upper-body motor function which is crucial to an individual's ability to function independently and general quality of life. Thus, exploration into assessments, such as the C3t, which may help inform us about an individual's upper-body motor function is vital, particularly for diseases which impact upper-body function, such as HD.

1.5.2.3.2.2 The Clinch Token Transfer Test

1.5.2.3.2.2.1 C3t: Overview

The Clinch Token Transfer Test was originally developed at Cardiff University by Dr Susanne Clinch during her PhD studies and was originally known under the moniker the Money Box Test (MBT). The test was originally developed due to a lack of objective clinical assessments in HD for functional capacity but was found to be related to numerous other areas of HD as well as discriminating between manifest HD and healthy controls (Bennasar *et al.*, 2018). Typically, the entire test protocol takes around 10-15 minutes to complete. The full manual for the C3t will be sent along with this thesis for examination.

The C3t is formed of six individual tasks taken in the following order:

1. The Baseline Transfer Task (BTT)
2. The Baseline Value Task (BVT)
3. The Complex Value Task (CVT)
4. The Baseline Alphabet Task (BAT)
5. The Complex Transfer Task (CTT)
6. The Dual Transfer Task (DTT)

The tasks can be classified as either ‘transfer tasks’ (BTT, CTT, DTT) or ‘baseline tasks’ (BVT, CVT, BAT). The time taken to complete each task is the primary measure recorded during the C3t.

1.5.2.3.2.2.2 C3t: Transfer Tasks

The BTT, CTT and DTT all require a series of 8 tokens of unique sizes to be transferred from their starting position on one side of the C3t board into a slotted box on the other side of the board. The tokens are transferred one at a time being picked up with the participants non-dominant hand, transferred into their dominant hand, and finally placed into the slotted box before the next token is picked up. Tokens are positioned on the board in order of physical size, with the largest token being the furthest from the participant and the smallest being closest.

Where the transfer tasks differ from each other are the extra rules that must be followed during each task.

The BTT requires tokens be transferred in order of physical size, starting with the largest and working down to the smallest.

The CTT uses a second set of tokens with numbers printed on them (each number if unique) and has the participant transfer the tokens based on the number printed on each, starting with the token with largest number and working down to the token with the smallest number.

The DTT uses a third set of tokens with numbers printed on them and has participants again transfer them in order of the numbers printed on the tokens (largest to smallest). Additionally, the DTT has participants say each letter in their native tongues alphabet, up to three times, whilst completing the task.

The rationale behind the increased cognitive loads of the CTT and DTT was that the presence of cognitive deficits in a participant would, in the presence of a cognitive load, exacerbate subtle motor symptoms. The exacerbated motor symptoms should then in turn increase the time participants take to complete the CTT and DTT, and so increase those tasks sensitivity to motor symptoms relative to the BTT, which has no such cognitive load.

1.5.2.3.2.2.3 C3t: Baseline Tasks

The baseline tasks (BVT, CVT and BAT) are included in the C3t to provide assurance that the participant is capable of counting and saying the alphabet when their full attentional capacity is given to the task.

During the BVT and CVT participants are presented with cards on which are listed the same numbers as on the tokens of the CTT, and DTT respectively. Participants are asked to read out the numbers on the cards starting with the largest number and working down to the lowest. During the BAT participants are asked to recite one in full the alphabet of their native tongue.

1.5.2.3.2.2.4 C3t: Recorded Measures

As mentioned, during all six tasks the primary measure recorded is the time it takes participants to complete each task. Additionally, there are 31 other scores listed in the C3t manual, which may be computed which can be broadly split up into recorded measures and derived measures. The distinction between recorded and derived measures is that recorded measures are directly observed during the C3t tasks whereas derived measures are calculated after the C3t has been completed from the recorded measures. Table 10 lists all C3t measures presented in the manual along with which tasks they apply to.

Table 10: Recorded and derived measures of the C3t

Recorded Measures		
Measure	Description/Equation	Applicable Tasks
Time taken (seconds)		All tasks (6)
Rule errors		All transfer tasks (3)
Transfer errors		All transfer tasks (3)
Dropped tokens		All transfer tasks (3)
Correct values		BVT, CVT (2)
Correct letters		BAT, DTT (2)
DTT alphabet time		DTT (1)
Total recorded measures: 20		
Derived Measures		

Measure	Description/Equation	Applicable Tasks
Transfer task accuracy		All transfer tasks (3)
Transfer task total score		All transfer tasks (3)
Value accuracy		BVT, CVT (2)
Alphabet accuracy		BAT, DTT (2)
Correct letters per second		BAT, DTT (2)
BTT-CTT time cost		Fusion of BTT & CTT (1)
BTT-CTT total score cost		Fusion of BTT & CTT (1)
CTT-DTT time cost		Fusion of CTT & DTT (1)
CTT-DTT total score cost		Fusion of CTT & DTT (1)
BAT-DTT alphabet cost		Fusion of BAT & DTT (1)
Total derived measures: 17		

1.5.2.3.2.2.5 C3t: Summary

The C3t is a novel clinical research assessment tool specifically designed for HD known to be related to numerous HD symptom domains. It should be noted that the C3t in its current incarnation is a research assessment tool, designed for clinical research only, rather than as a tool to inform clinical decision making. The test meets the requirements of being simple to use and cost effective (~£100/unit) as specified in section 1.5.2.2. Its nature as a dexterity-dependent transfer task makes it ideal for instrumentation as by combining it with simple sensors (discussed in the next section) a range of motor symptoms may be detectable. Given the high degree of similarity between the transfer tasks however it is plausible that not all will be needed in the final instrumented version of the C3t. This concept is explored in chapter 2 and chapter 4.

A significant downside to the C3t in its current state however is the large number of measures which must be manually derived by the researcher although the practical value of including every measure is unknown. The practical value of each recorded and derived measure should be explored with a view to simplifying the C3t by determining which measures should be retained and which can be discarded. This is covered more in Part 4: Chapter summary and thesis objectives.

1.5.2.3.3 Sensors

1.5.2.3.3.1 General Considerations

A crucial step in developing an instrumented assessment is deciding what sensors are suitable for instrumentation. With respect to this project there are three primary considerations listed below in rough order of priority.

1. Suitability of the sensors for detecting & assessing HD motor symptoms
2. The cost of the sensors
3. The ease of use of the sensors

Suitability is naturally the most important consideration – to be fit for purpose the instrumented assessment must have sensors suitable for assessing the motor symptoms of interest.

As stated in section 1.5.2.2, the ease of use of and cost of the instrumented assessment are important to consider as they directly impact is wide-scale adoption. This means that highly expensive sensors and/or those which require a great deal of specialised training to effectively operate should be avoided where possible. It should be noted that reducing costs where it is possible to do so is also an ethical issue, research projects are primarily funded in the UK by government organisations via taxpayers and charitable donations. As such, there is moral obligation for researchers to be mindful of the costs their projects incur – every pound that goes into one project is one that cannot be used for another. The ethical implications of sensor cost is of course less of a concern for commercial companies and would in that setting be more about efficiency.

1.5.2.3.3.2 Available sensors

There are a wide range of sensors that suitable for measuring human body motion. These range from more obvious solutions such as a simple tri-axial accelerometer, to more complicated inertial measurement units (IMUs) and electromyography (EMG) sensors to the gold-standard of movement analysis, full-body motion capture. Examples of these sensors are shown in Figure 1.

Figure 1: From left to right - GeneActiv accelerometer, Xsens IMU, Delsys EMG

Each sensor type has its own advantages and disadvantages which must be considered.

Tri-axial accelerometers whilst inexpensive and simple to operate collect only simple acceleration signals along the x-, y- and z-axes. Although they are widely used in many areas as shown in section, 1.4.3 there are much more sophisticated options available.

IMUs include accelerometers alongside gyroscopes and (sometimes) magnetometers, allowing for additional measures such as velocity, pitch, roll, yaw, and angular velocity to be derived. These devices range from moderately expensive to highly expensive and, in the case of full-body IMUs, often require the use of specialised software packages to synchronise multiple sensors and to provide reference locations for the calculation of displacement and joint angles (Filippeschi *et al.*, 2017).

Marker-based motion capture is currently the gold-standard of motion capture (Filippeschi *et al.*, 2017). Using an array of infrared cameras and a series of reflective markers attached to the participant in pre-specified locations highly accurate recording of motion can be achieved. The downside however is that such systems are highly specialised, extremely expensive, and often require long periods of setup. Marker-less motion capture systems are becoming available however the technology is still in its infancy as well as being expensive and again requiring a high degree of competency to properly utilise.

EMG sensors measure the subtle electrical signals produced by muscles as they contract and could be particularly useful for noticing the random contractions that will occur as a result of chorea. They can also be combined with IMUs allowing for multiple types of data collection simultaneously. EMGs suffer however from a high degree of variability between different subjects, particularly of differing body compositions (Trigili *et al.*, 2019; Lanza *et al.*, 2020). Whilst this drawback can be overcome via calibration and expert use, it makes the devices significantly more complicated to use relative to accelerometers or IMUs. This is particularly important if eventually the assessment might be placed into the home. It should be noted however that work is being undertaken to simplify & streamline EMG capture & processing, with one particular motivation being due to their potential use for exoskeletons and advanced prosthesis (Trigili *et al.*, 2019) .

1.5.2.3.3.3 Accelerometers & IMUs

The decision was made at the inception of this project to use two wrist-worn accelerometers along with a full-body set of IMUs. The concept was that in the event the accelerometers alone were insufficient for assessing motor symptoms the IMUs could provide additional data. A second phase of the project could then be undertaken to reduce the required number of IMUs from a full body set to an optimised smaller set to aid clinical adoption. The C3t is well suited to being combined with such sensors, the C3t transfer tasks providing a good opportunity to record movement data in a controlled manner. Whilst a full-body motion capture would have been more accurate, the difficulties in applying such technology to an HD cohort along with its reliance on expensive equipment often unavailable in clinics were felt to make it unsuitable for this project. Additionally, as discussed in section 1.4.3, there is a wide body of research that has successfully used accelerometers and IMUs to assess general & specific motor symptoms in PD and HD, setting a precedent for their selection.

The opportunity arose part of the way through the project to have multiple additional study sites collecting C3t data. These sites were provided with accelerometers rather than full-body suits of IMUs. As a result, given the significantly greater amount of data available with the C3t and accelerometers, the analyses in this thesis did not ultimately include IMU data. The limitations of this decision are discussed in 5. The accelerometers used for this project were GeneActiv tri-axis accelerometers (Activinsights; UK).

1.5.2.3.4 Sensor Features

1.5.2.3.4.1 General considerations

Instrumented assessments are assessments that use sensors or other electronic technology to take measurements during an assessment. The primary reason for using instrumented assessment over other types of assessments are measurements recorded by the sensors. Such measurements are known by many names but in machine learning and general multivariate analysis literature they are routinely referred to as features.

As has been previously stated so far there are two primary considerations regarding features with respect to assessing motor symptoms – their efficacy in detecting & assessing the motor symptom under investigation, and the ease with which they can be given clinical meaning.

The remainder of this section discusses general aspects of features and the process of handling acceleration data captured during the C3t for the purpose of assessing motor symptoms in HD.

It is important to note that that the explanations given are in the context of machine learning. The limited sample sizes available to this project prohibited the use of more advanced machine learning technologies (e.g., neural networks, clustering algorithms, etc). However, various types of multivariate regression analysis are used in the analyses presented in chapter 2 and chapter 4 and standard machine learning techniques (e.g., feature importance, cross-validation, etc) are employed to help ensure robust results. As such, the following discussion about features is presented in the context of machine learning.

1.5.2.3.4.2 What is a feature?

Features can formally be thought of as any discrete numerical quantity that in some way describes all or part of a phenomenon. Almost anything can be turned into a feature. Obvious examples include the mean and standard deviation of a series of data but more abstract concepts like colours can be used as well as long as they are somehow encoded numerically.

Throughout the machine learning literature and general multivariate analysis literature there are a number of common terms used involving features that will be referred to throughout this thesis.

Feature extraction refers to the process of directly extracting features from a set of data. *Feature engineering* can be thought of as a level up from feature extraction, often requiring multiple steps, domain knowledge and/or combining different types of data together. As an example, feature extraction might involve calculating the mean from a data series whereas feature engineering might involve using domain knowledge to group those means into different categories. *Feature selection* is used to describe the process of selecting the best features for a given task from a larger set of features. Finally, *feature importance* is the term used to describe how useful a given feature is estimated to be to the model it is used in.

When developing an instrumented assessment, the features extracted/engineered from the assessment measures will be one of the key properties that contributes to its success or failure. There are many different sensors that can be used to take similar measurements and numerous types of statistical models & tests that can be run to analyse the data, but if the right features are not utilised for the task at hand, then no amount of complexity or rigour elsewhere will make up for it.

This thesis is primarily concerned with turning acceleration signals captured during the C3t into features that are related to movement disorders in HD. As such, before thinking about what features might be suitable for use, the acceleration signals themselves should first be considered.

1.5.2.3.4.3 Accelerometers, acceleration, and their features

Acceleration is the second derivative of position with respect to time and the first derivative of velocity with respect to time, describing the rate at which velocity changes. The derivative of acceleration, jerk (i.e., the third derivative of position), has been used in a variety of fields including having been used to extract features from acceleration signals for use in instrumented assessments of motor symptoms in PD, particularly for postural instability (Eager, Pendrill and Reistad, 2016; Rovini, Maremmani and Cavallo, 2017). The fourth, fifth and sixth derivatives of position are snap, crackle and pop (Eager, Pendrill and Reistad, 2016) and are rarely used outside of physics.

Acceleration is typically captured by attaching a piezoelectric material (a material which will emit an electric charge in response to mechanical stress) of known mass to a static structure. A change in motion will exert force on the material causing a charge to be emitted and, as the mass is constant, the charge will be proportional to the force (i.e., Newtons 2nd Law) - acceleration. The sensor will be sensitive to acceleration along the axis it is ‘pointing’. A tri-axial accelerometer has three such sets of material facing different directions, capturing acceleration along x-, y- and z-axes, respectively.

Accelerometers only record allow acceleration, whereas more advanced IMU sensors combine accelerometers with gyroscopes and magnetometers allowing position, velocity, and orientation to be calculated. Although technically acceleration can be integrated to estimate velocity, without knowing the orientation of the sensor the error of the calculation will increase over time due to it being unknown which direction the sensor is facing and so how much of the acceleration is due to gravity. Jerk however can be calculated directly from acceleration data for a given time period.

The developed instrumented assessment was based around participants wearing two tri-axial accelerometers whilst they took the C3t, resulting in 6 signals being generated per C3t task and 18 signals per C3t test instance. These acceleration signals are of course by themselves are not sufficient to assess movement disorders and need to have features extracted from them in order to do so.

Features extracted from acceleration signals can be grouped into several domains – time, frequency, time-frequency and sparse (Krishnan and Athavale, 2018).

Time domain features are related to how a signals values change over time (e.g., mean, variance, entropy measures, etc). *Time-frequency* features describe signals in terms of both time and frequency, one of the main benefits being those small structures hidden by other larger structures within the signal can be revealed. *Frequency domain* features describe how a signal changes with respect to its spectral composition (e.g., spectral edge frequencies, dominant frequencies, etc). *Sparse domain* features further expand on time-frequency domain features by further decomposing the dynamic time-series signals.

As is stated in section 1.5.2.2, an aim of this thesis was to develop an instrumented assessment which relies on simple, tailored features that are easy to translate into clinical practice. This is in-line with the recommendation given in (Krishnan and Athavale, 2018) which suggests that first and foremost feature extraction should be based around application rather than just generating features each time data is analysed.

Frequency, time-frequency, and sparse domain features have been shown to be useful in a variety of context, including the assessment of movement disorders (Krishnan and Athavale, 2018). However, features produced using data from these domains are often harder to give clinical meaning to than features produced from time domain data. Additionally, there is significant precedent for using time-domain features for assessing motor symptoms in both HD and PD as discussed in section 1.4.3. As such, throughout this thesis there is a particular focus on features derived from the time domain.

There are numerous examples of time domain features that can be extracted from acceleration data throughout the literature. These range from simple measures like the mean, standard deviation, variance, and axis correlations to measures of structural properties such as signal peaks and widths to more complex measures like sample entropy and Lyapunov exponents.

Notably any feature that can be extracted from acceleration signals can also be extracted from its derivatives. As jerk, the first derivative of acceleration, has previously been shown to be effective for assessing symptoms in PD it will be explored here as well.

Once the instrumented assessment is complete with the underlying assessment having been taken, sensor data downloaded, and relevant features extracted the final step is explore the features relationship with the appropriate the gold-standard measures.

1.5.2.4 Instrumented Assessment Design: Showing Validity

Prior to discussing how validity of the instrumented assessment can be shown it is important to first discuss the concept of validity. There are multiple types of validity used throughout medical literature, two particularly notable types being content, predictive, criterion, and construct validity (Adams *et al.*, 2014; Bellamy, 2015).

Criterion validity estimates the extent to which an assessment agrees with some gold standard or otherwise absolute measurement. Construct validity focuses on whether an assessment is related to the construct/phenomena in question. Two further aspects validity, concurrent validity and convergent validity are also worth mentioning here. Concurrent validity, an aspect of criterion validity, is concerned with the level of agreement between two assessments, one of which is known to be the true value of a phenomena (Adams *et al.*, 2014). Convergent validity, an aspect of construct validity, refers to how closely an assessment is related to existing measures of the same construct (Krabbe, 2017).

Concurrent and convergent validity both appear at first to be very similar, however there is a subtle difference in terms of the question each seeks to answer, and the underlying assumptions made. Concurrent validity, as an aspect of criterion validity, assumes one of the two measurements measures the true value of a phenomena, and seeks to answer the question *"does this new measurement measure this phenomena?".* Convergent validity, as an aspect of construct validity, assumes one of the two measurements assesses (but is not necessarily the true value of) the phenomena of interest, and seeks to answer the question *"to what extent does this new measurement agree with this old measurement of the same phenomena?"*.

The key difference is concurrent validity assumes one of the measurements is the true value of the phenomena of interest, whereas convergent validity only assumes one of the measurements is related to the phenomena of interest (albeit ideally very strongly).

Ideally, we would focus on concurrent validity, however in this thesis we focus on evaluating the convergent validity of the instrumented and non-instrumented C3t (the new scale/measure) with existing measures of HD disease state (e.g., the UHDRS-TMS). The reason for this is that, as mentioned in the previous sections, whilst the existing measures of HD disease state are considered to be gold-standard assessments they are not accepted to represent the absolute 'true' value. It should be noted that this is a common issue with criterion validity for questionnaire-style clinical outcome measures like the UHDRS (Bellamy, 2015). Using chorea as an example, the chorea subset of the UHDRS motor assessment is not accepted to represent the true underlying level of chorea seen in a patient (Reilmann *et al.*, 2011a). We do however accept that chorea subset of the UHDRS motor assessment (along with the other existing gold-standards of HD assessment) are related to the underlying symptom phenomena. Thus, we set out to explore the convergent validity of the C3t with these existing measurements, and so explore whether the instrumented and non-instrumented C3t can be considered related to these phenomena.

This tactic of showing convergent validity with existing gold-standards, however imperfect, can be seen numerous times in section 1.4.3. Almost all the reviewed work in the literature on instrumented assessment for motor symptoms in some way provided evidence of their validity by tying the measures they produced to the relevant current gold-standard.

The actual process of showing the validity of an instrumented assessment (in this case the instrumented C3t) is conceptually straightforward and can be broken down into 3 core parts.

First, whether or not there is a relationship between the instrumented assessment and gold-standard measures in cross-sectional data needs to be determined (a.k.a., concurrent validity). Typical methods to do this common throughout scientific literature include visualisation (e.g., scatter plots, boxplots) and hypothesis testing (e.g., correlations & group differences). More advanced models can also be applied (e.g., regression models, classifiers, etc) along with machine learning techniques (e.g., cross-validation, feature importance, etc) to improve the robustness and simplicity of said models.

If a relationship is observed in cross-sectional data, the next step is to determine how the instrumented assessment performs in longitudinal studies. Ideally, the outputs metrics of the instrumented assessment should to some degree mirror any changes over time in disease symptom severity. Additionally, how robust the outcome of the assessment is, such as whether (and to what degree) the assessment is subject to a test-retest effect should be determined. This is vital as such effects, if any, will need to be accounted for before the assessment can be reliably said to be sensitive to change over time. The methods to do this again involve standard statistics such as group difference hypothesis testing as well as tests such as effect sizes and the analysis of distributions (upon which many effect size estimates are based).

Finally, it is necessary to assess is whether the instrumented assessment is clinically accepted, i.e., whether it is easy to use such that it can be widely deployed and whether the produced measures are accepted by the clinical community. Unlike the previous two points there is no straightforward method to judge this. Methods that might be used are clinical opinions based on interviews & questionnaires about the assessments and summary statistics on collected data quality.

It is important to note that this thesis focuses on the initial development of the C3t as an instrumented assessment of specific HD motor symptoms. As such, only the first stage of showing validity in cross-sectional data is explored in this thesis, with the remaining two steps being outside of its scope.

1.5.3 Designing a Remote Data Collection Platform

1.5.3.1 RDCP Design: General Overview & Rationale

As was mentioned in section 1.5.1, the opportunity arose during this project to vastly increase available data sample size by embedding the C3t and accelerometers into additional ongoing projects across disparate study sites. Collecting data, in particular non-standard sensor data, across multiple clinical study sties necessitates the construction of some sort of data collection platform which we have termed here an RDCP.

An RDCP was needed in this project for two main reasons.

Firstly, it can be used to facilitate & simplify the collection, transmission, and storage of C3t data (i.e., the traditional clinical assessment version of the C3t which records, among other measures, the time taken to complete tasks). As will be discussed in chapter 2, the C3t's large amount of recorded and calculated measurements make it unnecessarily laborious to record manually. This problem can be solved by collecting the C3t using a dedicated software collection system, allowing recorded measurements to be easily entered and calculated measurements automatically generated. Similarly, the large number of measurements and disparate study sites paper-based collection problematic. Whilst electronic files (e.g., excel) could be created and sent for collation & analysis, a simpler solution is to link the software collection system to a database, allowing for data to be stored in a single location ready for analysis.

Secondly, the inclusion of electronic sensor data during the C3t necessitates synchronising recorded sensor signals with the time C3t instances were taken. This can of course be accomplished manually using physical clocks and a stopwatch. However, a more elegant and accurate solution is to synchronise the internal clocks of the accelerometers and C3t software collection system. This a solution has the main benefits of being both less demanding of the user (simplifying clinical uptake), and more accurate by reducing the impact of human error reaction times.

As a result, the developed RDCP consisted of two distinct but connected systems – a C3t app used to take the C3t & synchronise with the accelerometer clock times, and a database backend used to store the collected data such that it can be retrieved and analysed later.

The specific design and architecture used for the systems is discussed in chapter 1, what is presented here are the general design considerations, decisions, and background information on the relevant software engineering principles.

1.5.3.2 RCDP Design: C3t App

1.5.3.2.1 C3t App: Functionality & Requirements

Based on the rationale presented section 1.5.3.1, in the C3t app should have four basic functionalities:

1) Record an instance of the C3t
2) Tie one or more C3t tests to a specific participant
3) Synchronise the C3t task timestamps with the sensor recording timestamps
4) Transmit recorded data to a remote database for storage

Each of these functionalities have specific requirements that should be met for the app to be considered operational.

The app should replicate as closely as possible the test procedure of the C3t. C3t task order should be enforced, and all measurements taken during the C3t should be possible to enter into the app. Each instance of a test should be related to a single participant, and as such the app will need to be capable of entering participant details. As the C3t is still under development, the app itself should be designed in such a way that specific functionality (e.g., tasks, recorded measures, derived measurements) can be added and removed. The app should be capable of automatically calculating of all derived C3t scores. Good software design principles should be followed, specifically, object-oriented programming (OOP) (Oriented, Programming and Oo, 2001).

As the app will be used alongside sensor technology, the app will need to able to synchronise its internal clock with internal clock of the sensors. This is needed so that individual C3t instances & tasks can be linked with recorded accelerometer data and ensure recorded sensor data is not included from either before the start or after the end of a C3t instance or task. This should be possible regardless of geographic location. Finally, the app will need to transmit recorded data, including participant details, to a remote database for storage and later analysis, and will need to do so securely.

As the C3t app is the user-facing portion of the RDCP an additional concern should be ease of use. The above functionality & requirements above should require as little training and be as self-explanatory as possible.

1.5.3.2.2 C3t App: Design Considerations & Principles

1.5.3.2.2.1 Development platform: Android vs. IOS vs. Microsoft Windows vs. macOS

The main design choice for the C3t app is which platform to develop it for. There are arguably four options - android tablet/smartphones (Google), IOS tablet/smartphones (Apple), Windows PC or Apple PC (macOS). The initial decision is whether to develop the app for a PC or a mobile (i.e., tablet/smartphone).

The benefit of developing for a PC, whether it be Windows or Apple, is the increased processing power and the lack of need to synchronise timestamps with the sensors (as GeneActiv accelerometers are configured using a PC). The downside however is that in research and clinical settings, PCs can be heavily regulated by internal IT departments and not all clinics will have immediate on-demand access to a PC they can install software on. Whilst a PC is of course required to operate the sensors, it should not be a requirement to use the C3t in clinical studies (as not all studies will have the sensors available or necessarily want to use them).

The benefit of using a mobile platform however is that (at time of writing) almost everyone has access in some form to a smartphone. Smartphone apps are also much better suited to the simple, singular purpose the C3t app would need to serve. Smartphones/tablets (android) are also relatively cheap in comparison to PCs, allowing them to potentially be supplied to clinical sites if needed. The downside however is that there is a large range of operating system versions even for the same base platform which require slightly different code to perform the same task. Additionally, as new operating system versions are released the app will require updating. Finally, if the sensors are in use then they will need to be configured on a PC and so the timestamps on the app will need to be synchronised to the PC.

1.5.3.2.2.2 Object oriented programming (OOP)

OOP is a programming paradigm widely used in application development. Whilst it is not a design consideration per-se, it is the foundation of the Java programming language which is used to develop the C3t app. Whilst a complete discussion of OOP is outside the scope of this thesis, several advanced ideas of OOP are made use of in chapter 3 and as such the core concept of OOP and the rationale for using it is introduced here.

Most modern programming languages have two fundamental concepts - data structures and functions.

- *Data structures* range from simple individual *bits* which can take the values 0 or 1, to more complex concepts like integers and floats (i.e., decimal points) to full data structures with their own internal structure and logic (e.g., linked lists, arrays, hash maps, etc).
- *Functions* (often called methods in OOP) are repeated procedures analogous to mathematical functions. A typical function will consist of a header and a body. The header contains arguments (i.e., variables) which the body uses in its execution. For example, a function *calculate_sum* could take two variables, *x* and *y* and return their sum.

OOP is essentially the fusion of one or more data structures (commonly referred to in OOP as attributes) with zero or more functions to create an object. Objects are used in many ways, but one common usage is to conceptualise real world concepts.

As is the case in many aspects of programming an example is more useful than the theory. A classic example routinely used in entry-level programming courses is a *BankAccount* object. A simple specification for such an object is shown in Table 11.

Table 11: BankAccount object example

BankAccount Object		
Attributes		
Name	Type	Description
account_id	integer	Unique identifier for the account
account_balance	float	Current account balance
Methods		
Name	Arguments	Description
make_payment	*receiving_account_id (integer)* *amount_to_pay (float)*	Send funds to an account
add_funds	*amount_to_add (float)*	Add fund to this account

Once an object is defined one or more instances of it can be created. Using the BankAccount example, a given individual might have one or more bank accounts which in the banks software system would mean that they have multiple distinct BankAccount instances.

The primary reason OOP is used in programming is because of how easy it makes programming the logic for otherwise abstract real-world concepts.

Using OOP, we can immediately recognise that the C3t app will need at a basic level two main objects; a Participant object which conceptualises the concept of an individual participant, and a C3t object which conceptualises the idea of an instance of the C3t. The full design specification of the C3t app based around OOP is given in chapter 3 along with descriptions of the more advanced OOP concepts that are made use of (e.g., inheritance, abstract objects, interfaces, etc).

1.5.3.2.2.3 Software development methodologies

Just as it is important to consider what technology is appropriate for use, it is important to consider what development methodology is the best suited when building software. Broadly speaking, there are two core software development methodologies - Waterfall and Agile (Palmquist *et al.*, 2013).

Waterfall, also known as the 'traditional' style of software development, is based on the following seven step sequential process.

1. System & software requirements
2. Analysis
3. Design
4. Coding
5. Testing & integration
6. Operations

In waterfall development there is first an initial requirement gathering phase during which the requirements of both the complete system and the software that it consists of are scoped out (i.e., what the system needs to do as a whole and what components will be required as a result). Second, the requirements are analysed so that the technical specifications of individual components can be drawn up (i.e., how the individual components will be constructed). Third, the complete system architecture is designed (i.e., how the components will fit together). Fourth, the coding of the components is performed (i.e., the software is written). Fifth, the components are tested and (if operational) integrated into the complete system (i.e., the software is tested, and the components joined together to form the system). Finally, the system is installed/deployed, and its functioning monitored with maintenance being completed as needed.

Unfortunately, whilst conceptually simple, Waterfall has been widely criticised due to the rigid structure it imposes on the software development lifecycle due to its sequential nature (Palmquist *et al.*, 2013). These criticisms include Waterfalls often naïve assumption that priorities will not change (i.e., that the requirements gathering process is both complete and correct), that no user authentication is done until after development has been completed, and that testing is conducted after every aspect of the coding has been completed (making fixing bugs very costly when multiple components must fit together).

As a reaction to the flaws of the Waterfall model a new development methodology, which is actually more of a philosophy than a specific methodology, was developed; Agile.

The Agile methodology/philosophy of software development was produced in 2011 by seventeen professionals who developed and agreed upon a set of values and principles that they felt should guide software development (Beck *et al.*, 2001). The Agile manifesto is based around the following four statements.

1. Individuals and interactions over processes and tools
2. Working software over comprehensive documentation
3. Customer collaboration over contract negotiation
4. Responding to change over following a plan

These statements result in Agile development methodologies having the following four core points (adapted from Palmquist *et al.*, (2013)).

1. Requirement gathering is done up front, but it is assumed requirements will change over time
2. The development iterations of steps 2-6 of the Waterfall model occur *per component* rather than for the system as a whole
3. Stakeholder participation is encouraged for each of the component iterations
4. Documentation is developed, but only as required

The primary benefit of Agile is in its agility relative to Waterfall. Whereas Waterfall development breaks the development process down into a series of sequential steps to be completed one after the other for all components, Agile encourages many of these steps to be performed in parallel across components. This results in a more fluid development cycle which is less rigid than the Waterfall model and so less prone to expensive re-writes and re-specs as requirements change, bugs are discovered, or alterations are requested.

The Agile development methodology is arguably the dominant software development methodology today (Hohl *et al.*, 2018). However, as is noted by Hohl *et al.*, (2018), it should not be viewed as a *"silver bullet"* suitable for all projects. In this project, it was felt the development of the RDCP was not well matched with the Agile methodology and that Waterfall was preferable. The reason for this was that before the projects inception the requirements of the RDCP had already been fully defined – a system was needed to facilitate the collection of the C3t data across multiple study sites and link C3t instances with sensor recordings. As such, it was felt that the simple, structured approach offered by Waterfall (which can result in faster development) was preferable to the flexibility offered by Agile. The result of this decision is discussed in chapter 4.

1.5.3.3 RCDP Design: Database backend

1.5.3.3.1 Database backend: Functionality & requirements

Based on the rationale presented section 1.5.3.1, the database backend should have four basic functionalities:

1) Storing C3t and participant data
2) Maintaining good data quality
3) Providing basic operational functions to the C3t app
4) Receiving, transmitting & storing data securely

The primary requirement of the database backend is that it be suitable for storing C3t and participant data recorded using the C3t app. The way in which the database structures C3t and participant data will primarily depend on the type of database, relational or non-relational, that is chosen (discussed in the next section). The ACID principles (Haerder and Reuter, 1983) of good database design should be followed to ensure data quality is maintained. As the C3t app effectively acts as an interface to the database, the database backend should provide standard operational functionality to the app, commonly referred to as CRUD operations (Create, Read, Update, Delete) (Martin, 1983)). Finally, the database should facilitate secure communication between itself and the C3t app by encrypting data during transmission using the Hyper Text Transfer Protocol Secure (HTTPS) (Rescorla, 2000). The physical location of the database will need to be secure.

1.5.3.3.2 Database backend: Design considerations & principles

1.5.3.3.2.1 Database Models

The main design choice needed to construct a database is what kind of underlying structure the data will be given within the database. The primary choice is between a relational structure and a non-relational structure.

Relational Data Model

Relational databases are based on the relational data model proposed by (Codd, 1970). Their main feature is their highly structured nature based around four primary concepts.

- **Tables**, made up of rows and columns, store data that can be logically grouped together (e.g., participant demographics)
- **Columns**, denote types of data in the table/the attributes of a table (e.g., date of birth, gender, etc)
- **Rows**, store the data of individual instances/records (e.g., participant: 123, date of birth: 12/09/1854)
- **Unique identifiers** are assigned to each row, allowing individual rows to be identified relative to all other rows even if multiple rows have identical column values in a given table

Non-relational Data Models

Non-relational data models differ from the relational model by eschewing the tabular data structure the relational model is based around. Unlike the relational model there are numerous examples of non-relational data models that do not necessarily have any commonality between other than not following the relational model.

Three of the more common structures are document stores, key-value stores, and wide-column stores.

Document stores, such as MongoDB (Anand and Rao, 2016), are structured around the concept of documents. Each document has a standard internal structure, but specific instances can vary drastically in terms of which fields are assigned. Documents are regularly compared to objects from OOP (see Object oriented programming (OOP).

Key-value stores, such as Oracle NoSQL (Anand and Rao, 2016), structure data based around the concept hash-maps/hash-tables. Hash-maps use hash functions to transform keys into a numerical index which refers to the location in an array within which is stored the keys associated value. Key-value stores take this concept and apply it to a database, allowing data to be stored and retrieved using key-value pairs.

Wide-column stores, based on (Chang *et al.*, 2006), make use of tables, columns & rows allow for nested structures to be created, essentially acting as two-dimensional key-value stores.

Relational vs. Non-relational Data Models

As with many technology choices, the deciding factor is the use case in question. Relational databases are highly structured, which makes maintaining data consistency very easy. The relational structure is also very human readable, with logical associations between tables, columns and rows being easy to understand. Non-relational models meanwhile are much better suited to very large datasets containing data which may not conform or be well described by the relational model.

For this study and the RDCP, the relational model was chosen. The primary reason for this is that the data collected during the project is highly structured and its scope limited. All C3t instances contain the same values and all participants have the same demographic information requirements. As such, the relational model is appropriate as this data can be very simply represented using the relational model. It should be noted however that for more general medical studies, non-relational data models should be preferred. The nature of these studies, especially those with large numbers of branching 'pathways' makes them ill-suited to the rigid structure inherent to relational databases. A recent study on the benefits of relational and non-relational data models for medical data agrees with this assessment (Sánchez-De-Madariaga *et al.*, 2017).

1.5.3.3.2.2 Create, Read, Update, Delete: CRUD

Create, Read, Update, Delete, operations (CRUD) (Martin, 1983) are the standard operations necessary to interact with a database and should be implemented. The operations themselves are fairly self-explanatory.

- **Create** operations create new rows, for example adding a new participant
- **Read** operations read data from the database, for example reading all participants demographics
- **Update** operations modify existing rows, for example updating a participant's date of birth
- **Delete** operations remove rows, for example deleting a participant who has withdrawn from the study

Whilst these operations are conceptually simple in isolation the complexity in their development comes from a single database being accessed by multiple users simultaneously. As such the ACID principles were developed.

1.5.3.3.2.3 ACID: Atomicity, Consistency, Isolation, Durability

The ACID principles, Atomicity, Consistency, Isolation, Durability, were laid out by (Haerder and Reuter, 1983) as a series of guidelines for database transactions, (e.g., CRUD) to follow. Whilst they are not a design consideration, they are crucial to the proper design of the RDCP database and so the database backed was constructed with the ACID principles in mind.

1.5.3.3.2.4 HTTPS

All data transferred over the internet is susceptible to a variety of attacks, ranging from simple interception to the injection of malicious code. The Hypertext Transfer Protocol Secure, HTTPS, is an extension to HTTP which adds a layer of security by encrypting the data using Transport Layer Security (TLS).

Connections between computers over a network are based around the concepts of sockets, ports, and IP (internet protocol) addresses. Computers communicate with each other by establishing socket connections through particular ports and sending messages over the network using their IP addresses to identify each other. Using a phone call between two people as an analogy:

- Computers are the two people on the call
- The IP address is the telephone number
- The network is the phone network
- The port is the specific phone being used by each person
- The socket is call itself, the ongoing connection which once the call is ended will terminate

Using the same analogy, HTTPS solves the problem of someone tapping your phone line to listen in on the conversation or re-routing the connection such that you end up not talking to the person you thought you were calling.

HTTPS does this using two mechanisms. Firstly, it allows a computer to verify the identity of the server it is trying to communicate with. Secondly, it encrypts all data between the two whilst the data is in transmit.

To verify identity HTTPS certificates are issued by certificate authorities, such as the Internet Security Research Group which provides a free HTTPS certificate service (Aas *et al.*, 2019). The encryption method used by HTTPS is based on public/private key cryptography.

- A key is used to encrypt/lock data
- A public key can be shared with any computer and can only be used to encrypt data
- A private key is never shared and is the only way to decrypt data that has been encrypted using the public key

A server will have a public key that it shares with other computers and a private key that it does not share but uses to decrypt information encrypted using its public key. By using public/private key cryptography two computers can ensure that the data transmitted between them can only be read by the computers involved in the connection. To ensure the data sent to/from the database backend is secure, HTTPS will need to be used.

1.5.3.4 RDCP Design: Showing Validity

Just as with the instrumented assessment, the RDCP also needs to have its validity verified. There are two primary data sources that can be used to assess the validity of the RDCP – the amount of data correctly collected and feedback from end-users.

The amount of data correctly collected, and the quality of the collected data the most vital aspects of the system. This can be measured in two ways. First, by observing total amount of data that could have been collected against the quantity of usable data actually collected. Second, by looking at problems that may have occurred. Descriptive statistics should be sufficient to assess both of these areas, allowing trends to be identified and potentially spot any problem areas which need addressing.

Even if the amount of data collected is optimal, it is important to also understand the user experience, if users do not find the system easy to use it will ultimately not be adopted. As such assessing feedback from end-users is required to understand the systems efficacy and identify ways in which it may be improved in order to refine it and inform future studies. Such data can be collected in the form of user experience questionnaires and semi-structured interviews. A variety of methods could be used to assess such data however the most straightforward is again simple descriptive statistics and distribution analyses.

1.5.4 Summary

The development of a complete, functional RDCP is non-trivial. Collection systems are increasingly important in medical research as they allow unprecedented access to patients and clinicians whilst minimising the organisational overhead traditionally incurred during large studies. The first version of a system however is rarely the final one, multiple iterations of a system are to be expected in order to refine the user experience, remove any bugs or pain points in the software as well as update the capabilities as requirements change over time. It is however important to treat such systems as if they are the final version, designing them with both the clinical and technological needs in mind.

1.6 Part 4: Chapter summary and thesis objectives

The purpose of this chapter has been to set the stage for this thesis, showcasing the need for instrumented assessment in HD and detailing the various design considerations and technical underpinnings necessary to develop a final instrumented assessment capable of use in a clinical setting.

Overall, the aim of this research is to instrument the C3t, a clinical assessment known to be related to HD disease state, in such a way that it is suitable for assessing specific motor symptoms. To achieve this there are three high-level objectives.

Objective 1; further advance the current (non-instrumented) C3t by looking at its relationship with composite measures of HD disease state, specific motor symptoms, and removing unnecessary measures & derived scores.

Objective 2; produce the two components of the RDCP, deploy them in a series of research studies, collect the resultant data and assess the systems validity.

Objective 3; use the data collected during this project as well as via the RDCP to assess the suitability of the C3t for assessing whole-body and upper-body chorea.

Chapters 2, 3, and 4 address objectives 1, 2, and 3 respectively. It should be noted that the development of the C3t covered during this thesis is for the C3t as a clinical assessment research tool, rather than as a clinical tool to aid clinical practice & management. The distinction is that clinical assessment research tool is specifically for use in research and is not designed (currently) to directly inform clinical management of the disease. The C3t is nonetheless a clinical assessment (i.e., an assessment to be performed in clinic) and so, as was mentioned at the beginning of this chapter, for brevity be referred to as a clinical assessment.

Chapter 2: Continued development of the C3t as a clinical research tool for Huntington's Disease

2.1 Chapter Overview

As established in chapter 1, there is a clear need for more sensitive assessments of motor symptoms in HD. Sensitively assessing motor symptoms can not only aid clinical trials and evaluations but also clinical management of HD. Composite measures which combine motor symptoms with other symptom areas (namely cognition and function) show increased sensitivity to disease progression but could benefit from more sensitive assessments of motor symptoms. Instrumented assessments have been shown to be effective for assessing motor symptoms in PD, but have not yet been widely studied in HD, particularly regarding specific motor symptoms rather than general motor dysfunction.

In order to assess the validity of an instrumented assessment it is useful to define a baseline metric its performance can be measured against. As was noted in in chapter 1, one of the objectives of this thesis is develop the C3t into an instrumented assessment sensitive to chorea in HD. Thus, the first objective of this chapter is to explore the relationship, if any, between the non-instrumented C3t and chorea.

Additionally, this chapter takes the opportunity presented to further develop understanding of the C3t by exploring its relationship with composite measures of HD disease state, its stability over short time periods, and whether it is anchored to progression of symptoms over long time periods.

Finally, as was repeatedly mentioned in chapter 1, an overarching aim of this project was to develop an instrumented assessment that is easy to translate into clinical practice. Whilst much of this effort is based around the signal features discussed in chapter 4, simplifying the non-instrumented C3t will also help achieve this aim. As such, both the C3t protocol (i.e., the tasks it contains) and C3t scoring (i.e., the measures recorded during it) are critically analysed here and suggested for either retention or removal.

Specifically, the objectives of this chapter are as follows.

Objective 1: Establish the non-instrumented C3ts scores relationships with chorea.

Objective 2: Establish the non-instrumented C3t scores relationships with composite measures of disease state and assess their sensitivity to change in disease state over time.

Objective 3: Refine both the C3t protocol and the C3t scoring by critically assessing the C3t tasks and scores, ultimately providing recommendations for their retention or removal.

This chapter is an extension of the analysis conducted and reported in Woodgate *et al.*, 2021.

2.2 Introduction

Currently no clinically proven disease modifying therapies exist for HD, although numerous potential therapies are under active investigation (McColgan and Tabrizi, 2018). Many of these therapies aim to slow or even entirely prevent neuronal tissue damage and so are often designed to target the earlier stages of HD before severe damage has occurred when they may have the greatest impact.

A standard technique for showing efficacy of a therapy designed to slow or stop the progression of a disease is to show the difference in rates of progression between a treated group and an untreated control group over the same time span. Whilst conceptually simple, the complexity lies in how one measures progression of the disease in a sensitive and objective manner that is suitable for regular, large-scale collection.

To achieve this in HD there are several options, including using imaging techniques to assess the volume degradation in the affected regions of the brain, monitoring the amount of the mutant huntingtin protein (which causes the degradation), and assessing the severity of the symptoms the degradation causes. Each of these have their own advantages and disadvantages.

Imaging techniques give a direct biological/physical estimate of the extent the disease has progressed. They are also however expensive to be applied regularly at scale both in terms of economic and time costs. Monitoring levels of mutant huntingtin protein is cheaper than imaging but invasive and is only relevant to therapies seeking to inhibit production of the protein. Assessing symptoms however is efficient both in terms of time and resources, making it particularly attractive for large-scale longitudinal studies as well as general clinical practice.

As such, although symptoms are surrogates for the underlying structural changes occurring in the brain, if they can be measured in a sufficiently sensitive manner, they can provide a sensitive estimation of disease progression. Additionally, as was discussed throughout chapter 1, they are vital for the proper clinical management of HD particularly when it comes to assessing an individual's quality of life. If the collection protocol is sufficiently simple, datasets can be collected at a regularity and scale impractical for other methods, allowing for more complex analyses to be conducted and higher statistical power.

HD symptomatology is however complex, diverse, and progressive making it difficult to sensitively measure. Symptoms in HD are typically labelled as one four types – motor disorders, cognitive deficits, behavioural abnormalities, or functional impairment (McColgan and Tabrizi, 2018). Each of these symptom domains has one or more gold-standard assessments routinely used by the clinical and research community to assess the severity of a patient's symptoms. Whilst the following provides a brief overview of the various gold-standard assessment batteries a deeper level of information and discussion can be found in section 1.3.3.

The gold standard assessment batteries for the motor, cognitive, and functional domains are listed in the UHDRS (Kieburtz *et al.*, 1996). The UHDRS does contain a behavioural assessment battery, however the PBA-s is usually preferred (McColgan and Tabrizi, 2018).

HD motor symptoms are assessed using the UHDRS-TMS, cognitive symptoms using the SDMT, the SCWT & LFT, and functional symptoms using the TFC, FAS & IS (Kieburtz *et al.*, 1996; Wild and Tabrizi, 2014; McColgan and Tabrizi, 2018; Mestre, Bachoud-Lévi, *et al.*, 2018).

Behavioural abnormalities are typically accepted to not be reliable indicators of disease progression in HD (McColgan and Tabrizi, 2018). Conflicting results have been reported for cognitive symptoms with one series of studies finding they progress reliably in early manifest HD over 12-, 24- and 36-months (Tabrizi *et al.*, 2011, 2012, 2013) but a second larger study with more frequent measurements reporting only erratic, small changes (Meyer *et al.*, 2012). Similarly, functional capacity was found to be unreliable in one set of studies (Tabrizi *et al.*, 2011, 2012, 2013) but reliable in another (Meyer *et al.*, 2012). Motor symptoms however were found to reliably change during early-stage HD in both studies (Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012). Crucially, none of the gold-standard measures of motor, cognitive or function were found to reliably detect changes in pre-manifest HD (Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012). Overall, it appears that out of the gold-standard assessments the UHDRS-TMS is typically the most reliable for detecting change over time (Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012).

Although the literature shows the UHDRS-TMS can reliably detect some level of change in early manifest HD, the assessments validity has still been questioned based on its perceived insensitivity and known floor effects in early HD (Reilmann *et al.*, 2011a). The UHDRS-TMS is calculated by summing a series of 31 ordinally rated (0-4) assessments of motor function each of which assesses the severity of a motor symptom in one or more area (Kieburtz *et al.*, 1996). Sub-items from the UHDRS-TMS relating to a single motor symptom (e.g., chorea) are also routinely summed to create a severity measure of that specific symptom. However it is thought that these ordinal scales are unlikely to be sensitive enough to measure small amounts of progression of individual symptoms nor of general motor dysfunction, limiting their use for assessing progression (Reilmann *et al.*, 2011a).

The lack of an accepted method for sensitivity assessing changes in early-stage disease symptoms has led to continued research into assessment measures for HD. One promising avenue that has seen active research over the last few years is composite measures. Composite measures seek to enhance the sensitivity of available measures not by improving the measures themselves but rather by combining them together. Given the inherently multifaceted nature of HD this concept has the clear benefit of considering HD symptoms all at once rather than in isolation from each other.

The CUHDRS (Schobel *et al.*, 2017) PI_{HD} (Long *et al.*, 2017) are recently developed composite measures designed for assessing HD. The CUHDRS combines the gold-standard measures of motor, cognitive and functional impairment to produce a single index that has been shown to have increased sensitivity to change in early manifest (but not pre-manifest) HD over 12-months relative to its component measures (Schobel *et al.*, 2017). Similarly, PI_{HD} combines measures of motor and cognitive function together with age and the number of cytosine, adenine, guanine (CAG) polyglutamine repeats (the gene expansion responsible for HD) to produce an index representative of future risk of motor diagnosis based on the levels of these variables (Long *et al.*, 2017). The normalised variant of PI_{HD}, PIN_{HD} (Prognostic Index Normalised), enhances interpretation by centring PI_{HD} around a 0.5 survival probability (Long *et al.*, 2017). Both PI_{HD} and PIN_{HD} can be used to provide information about estimated progression rates but are not in themselves measures of progression.

Composite measures offer improvements, both conceptually and practically, over measures of individual symptoms. However, relative to the other measures they are still in their infancy although they are the topic of active research. However, a limitation of these composite measures is that they are based on the existing gold-standards, which are themselves known to be insensitive. Therefore, the sensitivity of the composite measures could be further enhanced in sensitivity of underlying measures they rely on were also enhanced. For example, it is known that the UHDRS-TMS is insensitive to early motor symptom progression. As the UHDRS-TMS is the only measure of motor function fed into the CUHDRS, PI_{HD}, and PIN_{HD} these measures cannot be more sensitive to changes in motor function than the UHDRS-TMS is. Overall, despite composite measures being a clear improvement over existing gold-standard HD assessments for measuring progression, there is still a need to enhance the baseline assessments used for HD symptoms.

As clinical trials tend to target the very earliest stages of HD (i.e., pre-manifest, prodromal and early manifest) assessments are needed that are sensitive to the symptoms seen throughout these stages. It should be noted that cognitive and functional impairment does occur during these stages and are thought to often pre-date motor symptoms. However, motor symptoms still feature prominently during the early-stages of the disease and are used to define when HD has entered it’s manifest stage

(McColgan and Tabrizi, 2018). In particular, in terms of motor symptoms early-stage HD is typically characterised by hyperkinetic disturbances (e.g., chorea, tics) which as the disease progresses eventually give way to hypokinetic disturbances (e.g., bradykinesia, rigidity) (McColgan and Tabrizi, 2018). Therefore, one route to enhancing the sensitivity of composite measures to early-stage HD, as well as improving HD assessment in general, is to develop more sensitive assessments specific to these symptoms.

One promising avenue of research, as is discussed at length in section 1.4, is instrumented clinical assessments (a.k.a. instrumented assessments). Such assessments typically combine some sort of action task with modern sensor technology to produce types of data not recorded in traditional clinical assessments.

Instrumented assessments have been used to great effect in Parkinson's Disease for assessing both individual and general motor dysfunction (Rovini, Maremmani and Cavallo, 2017). Whilst similar work has been attempted in HD, the majority of these attempts focus on gait, postural, or the UHDRS-TMS rather than specific early-stage motor symptoms. Those that have focused on early stage motor symptoms have shown at best moderate links with the underlying symptoms (Reilmann and Schubert, 2017b).

A crucial part of developing an instrumented assessment is to assess its validity by studying its relationship (or rather the measurements it produces) with the gold-standards (i.e., the UHDRS-TMS and/or its sub-items in HD). As this thesis is concerned with developing the C3t into an instrumented assessment sensitive to chorea, this chapter seeks to establish whether there is a pre-existing relationship between chorea and the non-instrumented C3t.

As the C3t is still in its infancy as an assessment the opportunity is taken to conduct additional analysis of the non-instrumented C3t.

It is already known that C3t scores are related to multiple symptom domains including cognitive, motor and function (Clinch, 2017a). However, the C3t relationship with CUHDRS, PI_{HD} and PIN_{HD}, are as yet unknown although considered likely given the components of the component scores. If the C3t is closely linked to these composite measures then it could potentially be used as a surrogate for them, allowing estimations of them to be calculated regularly & rapidly without the training required by the measure's component UHDRS assessments. If the protocol is simple enough, it could even be potentially conducted in the home by a patient's family or primary carers, allowing for significantly more regular data collection than is currently practical. As such, the C3ts relationship with the CUHDRS, PI_{HD} and PIN_{HD} is explored here. Whilst the C3ts relationship with the UHDRS-TMS has been determined previously (Clinch, 2017a), a significantly larger sample size was available to this study and as such similar analysis conducted albeit using more sophisticated techniques (cross-validated regression models).

Analysis of the C3t so far has utilised cross-sectional data (Clinch, 2017a). It is unknown whether the C3t is 'anchored' to changes in the UHDRS measures (i.e., are changes in the UHDRS scores mirrored in changes to the C3t scores). Whether a clinical assessment is anchored to gold-standard measures is important as it provides evidence that the assessment is directly affected by the progression of the symptoms those assessments are known to assess. As a limited amount of longitudinal data was available to this study whether the C3t scores are anchored to the UDHRS scores was also explored.

Finally, the C3t produces a large number of scores not all of which are necessarily required for the C3t to be useful as a clinical assessment and many of which are very similar to each other, as are many of the C3t tasks. As such, there may be little practical benefit in collecting all of them. If the number of scores and/or tasks could be reduced the test itself would be simplified. This would help to ease clinical uptake of the non-instrumented C3t as well as the instrumented C3t (in the case of removing some of the C3t tasks). As such, the distributions and cross-correlation of the C3t scores are analysed with a view to providing recommendations as to whether each score (and ultimately their relevant task) should be removed or retained.

2.3 Methods

2.3.1 Data Collection

2.3.1.1 Participants

Data used in this study were drawn from four separate studies – PACE-HD, CAPIT-HD, TRIDENT and Developing Clinical Applications for a Novel Multi-Task Functional Assessment: The Clinch Token Transfer Test (referred to here as C3t PhD).

PACE-HD (Clinical trials registration: NCT03344601) is an ongoing multicentre intervention trial with sites in Europe and the USA where recruited participants were also participating in Enroll-HD. As PACE is an intervention study only baseline is included here prior to any intervention taking place.

CAPIT-HD was a multicentre European study (Cardiff and Manchester, UK; Créteil Paris, France; Muenster, Germany). In three sites (Cardiff, Manchester, and Muenster) participants were recruited from those participating in Enroll-HD. At one site (Créteil) participants were recruited from the ongoing Predictive Biomarkers for Huntington's disease study (Clinical trials registration: NCT01412125). CAPIT-HD provides the longitudinal dataset used in this study, with recruited participants completing a battery of assessments during a baseline visit before being asked to return for 1-month and 12-month follow up visits.

TRIDENT and C3t PhD are single-site studies in based in Cardiff (UK). All participants recruited were also enrolled in Enroll-HD and were recruited for a single baseline visit only. Notably, half the data from the C3t PhD study was collected by the author. Collecting and working with patients provided various benefits to this thesis as is touched on in chapter 4 and more broadly discussed in chapter 5.

Ethical approval for all studies was granted by Health and Care Research Wales (CAPIT-HD2 REC: 17/WA/0014, TRIDENT REC: 18/WA/0182, C3t REC: 17/WA/0014). All recruited participants had previously been genetically confirmed to carry the HD mutation, were 18 years or more of age and had the capacity to provide informed consent. Using the TFC and DCL participants were sub-divided into disease stages for each visit as shown in Table 12.

Table 12: TFC disease stage assignment requirements. DCL is Diagnostic Confidence Level and quantifies a clinician's opinion that any motor disturbances are due to HD. TFC stands for Total Functional Capacity score and quantifies the ability of a patient to perform daily life tasks. Both the DCL and TFC are fully described in section 1.3.3.6.

Disease Stage	Requirement	Broad description
Pre-manifest	DCL < 1	Yet to manifest overt motor symptoms
Prodromal	DCL = 2-3	
TFC Stage 1	DCL = 4; TFC = 13-11	Manifest HD, symptomatic
TFC Stage 2	DCL = 4; TFC = 10-7	
TFC Stage 3	DCL = 4; TFC = 4-6	

2.3.1.2 C3t scores & clinical scores

2.3.1.2.1 C3t Scores

All participants performed the C3t or its previous version, the Money Box Test (MBT), at a baseline visit with a subset also performing the same test at 1-month and 12-month follow up visits. The CAPIT-HD study included the MBT in its protocol whilst PACE-HD, TRIDENT and C3t PhD included the C3t. As such, there is a set of participants with C3t data and a set with MBT data.

The difference between the C3t and MBT are within the tasks each contains as indicated in Table 13 with shaded cells representing presence in that version of the test. It should be noted that tasks present in both the C3t and MBT are identical.

Table 13: Task presence in the C3t and MBT. The C3t contains two tasks that are not in the MBT – the DTT and CVT. The MBT contains one task that is not in the C3t – the TTT.

Task	In C3t?	In MBT?
Baseline Transfer Task (BTT)	Yes	Yes
Complex Transfer Task (CTT)	Yes	Yes
Dual Transfer Task (DTT)	Yes	No
Triple Transfer Task (TTT)	No	Yes
Baseline Value Task (BVT)	Yes	Yes
Complex Value Task (CVT)	Yes	No
Baseline Alphabet Task (BAT)	Yes	Yes

As this thesis focuses assessing movement symptoms only the transfer tasks (BTT, CTT, DTT & TTT) are of interest. The baseline tasks (BVT, CVT & BAT), unlike the transfer tasks, contain no movement component that should be affected by HD motor symptoms and so were not included in this study.

In terms of transfer tasks, the difference between the C3t and MBT and shown in Table 13 is the presence of the DTT in the C3t and the TTT in the MBT.

Both the DTT and TTT are the final transfer task in their respective tests and are identical in all one key detail – the numbers used on the TTT tokens are identical to those on the CTT tokens (although they are in a different order), whereas the DTT changes the numbers on tokens. Whilst a small change, this means the two tasks cannot be treated as the same. Data collected in the CAPIT-HD study used the MBT and the other studies the C3t, this means that the datasets cannot be fully merged, with a large group having performed the BTT and CTT and two smaller subsets having performed either the DTT or TTT. Due to the decreased sample size of the DTT and TTT they were omitted from this study with analysis focusing on only the BTT and CTT.

Table 14 shows the complete list of BTT & CTT scores extracted from the combined datasets used in the analysis. For simplicity, they are referred to collectively in this chapter from this point on as *the C3t scores*. Similarly, whilst data is drawn from the MBT and C3t they will be collectively referred to from hereon as *the C3t*.

Table 14: Extracted C3t scores used in the analysis presented in this chapter.

Recorded Measures		
Measure	Description/Equation	Number of measurements
Time taken (seconds)	The time a task took to complete in seconds (accurate to 2 decimal places)	2
Rule errors	The number of tokens picked up in the wrong order (maximum 8)	2
Transfer errors	The number of times participants did not correctly transfer tokens between their hands (maximum 8)	2
Dropped tokens	The number of times participants dropped tokens (maximum 8)	2
Total recorded measures: 8		
Derived Measures		
Measure	Description/Equation	Number of measurements
Transfer task accuracy	$\frac{16 - (rule\ errors + transfer\ errors)}{16} * 100$	2
Transfer task total score	$\frac{8 - dropped\ tokens}{time\ taken\ in\ seconds} * transfer\ task\ accuracy$	2
BTT-CTT time cost	$\frac{CTT\ time\ taken - BTT\ time\ taken}{CTT\ time\ taken}$	1
BTT-CTT total score cost	$\frac{CTT\ total\ score\ - BTT\ total\ score}{CTT\ total\ score}$	1
Total derived measures: 6		

2.3.1.2.2 Clinical Scores

In addition to performing the C3t, participants were also assessed using the full UHDRS assessment battery. Three measures of motor function from the UHDRS were used in this study – whole-body chorea, upper-body chorea, and the UHDRS-TMS. Whole-body chorea and upper-body chorea scores were calculated by summing chorea assessments of relevant areas from the UHDRS motor assessment as shown in Table 15. The single score used for the UHDRS-TMS was extracted as-is from the UHDRS motor assessment dataset.

Table 15: Chorea assessment area for the whole-body and upper-body chorea measures.

Chorea Assessment Area	Whole-body chorea	Upper-body chorea
Head	x	x
Face	x	x
Trunk	x	x
BOL	x	-
Left-upper limb	x	X
Right-upper limb	x	X
Left-lower limb	x	-
Right-lower limb	x	-

In addition to the measures of motor symptoms the CUHDRS and PIN_{HD} composite scores were also used in this study.

The CUHDRS was taken to represent general disease HD disease state and is calculated using the equation below by combining the TFC, UHDRS-TMS, SDMT and SWRT all of which were extracted from the UHDRS.

$$cUHDRS = \left[\left(\frac{TFC - 10.4}{1.9}\right) - \left(\frac{TMS - 29.7}{14.9}\right) + \left(\frac{SDMT - 28.4}{11.3}\right) + \left(\frac{SWR - 66.1}{20.1}\right)\right] + 10$$

PIN_{HD} was taken to represent projected disease progression and is calculated using a patient's UHDRS-TMS, SDMT, age and number of CAG repeats. PIN_{HD} is the normalised version of PI_{HD} relative to a 0.5 survival probability where $PIN_{HD} < 0$ indicates a greater than 50% chance of 10-year survival and $PIN_{HD} > 0$ indicates the opposite. This is done in the original paper in order to simplify explanation of the value and as such is preferred here for this same reason. PI_{HD} was not included in the analyses here as the relationship between PIN_{HD} and PI_{HD} is linear. PIN_{HD} is calculated using the following equation.

$$PIN_{HD} = \frac{PI_{HD} - 883}{1044}$$

Where:

$$PI_{HD} = 51 \times TMS + (-34) \times SDMT + 7 \times Age \times (CAG - 34)$$

Each clinical measure (whole-body chorea, upper-body chorea, UHDRS-TMS, CUHDRS, and PIN_{HD}) were extracted at the baseline visit. Where available, they were also extracted from the 12-month follow-up data. Where UHDRS data collected time-of-recording was not available, data collected within at most 6-months was used. PIN_{HD} was calculated only for participants in either the pre-manifest or prodromal stages of HD as the metric was designed for use in these populations rather than in manifest HD.

2.3.1.3 C3t App

As noted in chapter 1, a limitation of the C3t is the number of scores which need to be calculated. Even with a reduced number of C3t tasks analysed here, 8 scores must be captured throughout the test (4 per task) by the researcher and 6 further scores derived from them (14 total).

During CAPIT-HD, C3t data was collected on paper and the derived scores calculated either by hand, using excel macros or, in some cases, not at all. The difficulty of using paper-based collection methods will be discussed in chapter 3, but in short paper-based collection can lead to calculation errors and increases the level of work required by the clinician. Additionally, the eventual integration of the sensor data into the C3t requires more accurate timestamps be collected than what is possible simpler methods such as stop watches. As a result, during this project a C3t android app, was developed to aid clinical data collection. The C3t app and associated software systems are discussed fully in chapter 3.

All C3t data presented here from the PACE-HD, TRIDENT and C3t PhD studies were collected using the app and transmitted to a secure database held at the Centre for Trials Research (Cardiff University, School of Medicine, Cardiff, UK).

2.3.2 Data Analyses

2.3.2.1 Analyses-objective breakdown

The objectives stated in section 2.2 were addressed using a variety of methods each of which is described, in the order it was performed, in the following subsections.

2.3.2.2 Step 1: Histograms

Histograms were used to determine the distribution of each of the C3t scores. Whilst histograms are widely used for assessing normality, they can also be used to show whether variables are likely to contain useful information about the population they are drawn from. Invariant distributions can be used as a quick, reliable indicator that a variable is unlikely to contain useful information. Essentially, if almost all of a population under investigation records the same or very similar value for a measure it is unlikely that measure contains useful information regarding aspects of the population known to vary (like UHDRS scores). In this study histograms were used to remove C3t scores unlikely to contain useful information about the HD population. Visual inspection of histograms produced from each C3t scores were used to determine whether scores were sufficiently invariant to warrant being removed.

2.3.2.3 Step 2: Normality

The normality of each C3t score retained after Step 1 was assessed using histograms, Q-Q plots and three statistical tests (D'Agostino K-Squared, Anderson-Darling, and Shapiro-Wilks). As most scores were found to be non-normal, non-parametric data analysis methods were employed.

2.3.2.4 Step 3: Scatter plots

A surprisingly often overlooked but crucial assumption of correlation statistics and linear models is the presence of a monotonic relationship between the dependent and independent variables. If no monotonic relationship is present, then correlation statistics which depend on a monotonic relationship being present lose their meaning and linear model prediction accuracies cannot be trusted. A simple way of testing the assumption of a monotonic relationship is by using scatterplots (Schober and Schwarte, 2018). Whilst analysing scatter plots can prove impractical when working with very large amounts of variables this was not the case here with the analysis involving at most 84 pairs of independent-dependent variables.

Each C3t score retained after Step 1 was plotted against each of the six clinical measures. C3t scores judged from these plots to not show a monotonic relationship with *any* of the clinical measures were omitted from further analysis.

2.3.2.5 Step 4: Correlation & regression

To determine the relationship between retained C3t scores and the clinical scores correlation statistics and regression analysis were used. Correlation and regression analysis were performed only on C3t-clinical score pairings where monotonic relationships were thought to exist from Step 3. Note that only correlation analysis was performed for PIN_{HD}, as the sample size of the pre-manifest and prodromal population was deemed too low for regression analysis to provide meaningful results (n=16).

Correlation strength provided a quantification of the degree to which a monotonic relationship exists between the independent and dependent variables. Additionally, p-values provided by the correlation statistics were used to determine the likelihood that any observed relationship was just by random chance.

Regression analysis was used to determine how well each clinical score could be estimated by the retained C3t scores. Cross-validation (discussed below) was applied when constructing the regression models in order to reduce the likelihood that an overoptimistic estimation of model performance would be reported due to the model overfitting to the data.

Spearman's R was used in place of Pearson's R as the correlation statistic as the C3t scores were found to be non-normal. Holm-Bonferroni corrections (discussed below) were applied post-hoc *per dependent variable* to account for the multiple comparisons being made.

Standard linear regression was used for the CUHDRS (which is continuous) and ordinal linear regression was used for the clinical scores derived from the UHDRS motor assessment (all of which are ordinal). All regression models were cross-validated using a repeated k-fold cross validation strategy (k=4, repeats=10) and their estimation quality assessed using the Mean Absolute Error (MAE) and Normalised MAE (discussed below). Regarding regression model results, the reported MAE and Normalised MAE metrics are the mean and standard deviation of the MAE and normalised MAE across all cross-validation folds.

As data were drawn from multiple sites and two test versions, it was possible the sites and test versions would influence the C3t scores, which could lead to overoptimistic regression model performance. As such, a set of lasso regression models were built for each studied C3t score & clinical variable pair for which regression models were constructed, with the site each participant was recruited in included and the test version included as one-hot encoded independent variables. These models were cross-validated using the same strategy used for the ordinal linear regression models (k-fold cross-validation, k=4, repeats=10). The mean and standard deviation of each variable's coefficient was calculated per C3t score & clinical variable pair across all cross-validation folds, per model. The magnitude of these coefficients was then used to judge the effect of the sites and test versions on regression model quality, relative to the C3t scores.

Multiple hypothesis tests were run in this study (i.e., multiple Spearman's R correlations calculated) and post-hoc Holm-Bonferroni corrections were applied to minimise the likelihood of type I errors (i.e., the null hypothesis being rejected by chance). Holm-Bonferroni corrections differ slightly from traditional Bonferroni corrections however both have the same idea – the alpha value (the value below which p is considered to be statistically significant) is adjusted based on the number of observations made.

Traditional Bonferroni corrections adjust the alpha value by dividing the original significance level (e.g., 0.05) by the number of observations made. Bonferroni corrections are typically seen as overly conservative which has led to alternative measures being developed (Perneger, 1998). One such method is the Holm-Bonferroni correction which modifies the original Bonferroni correction procedure by adjusting the alpha level for each observation (Holm, 1979). The procedure of applying Holm-Bonferroni corrections is as follows.

1. *Gather each p-value for all observations and rank each in descending order (e.g., lowest p-value is given rank 1, second lowest rank 2 etc)*
2. *For each p-value (lowest to highest),*
    a. *calculate the corrected alpha level using the equation below*
    b. *if the current p-value is less than its corrected alpha value, then accept it as statistically significant and continue, otherwise reject it and all subsequent p-values as not statistically significant*

$$alpha = \frac{target\ alpha\ level}{n - current\ p\ value\ rank + 1}$$

2.3.2.5.1 Cross validation

Cross validation is a standard technique in machine learning to avoid overfitting models to the available training data and estimate the model's ability to handle unseen data (i.e., how general the model is).

Overfitting in machine learning refers to the issue of statistical models being over-optimized to for the training data in use. This over-optimisation can lead to overestimating the quality of trained models on training data and those same models then performing poorly on previously unseen 'real world' data (if it is not very similar to the training data).

Cross validation strategies are routinely employed to reduce the likelihood of overfitting models. One such strategy employed throughout this thesis is k-fold cross validation. During k-fold cross validation available data is split into 'k' evenly sized subsets (or 'folds'). Each subset takes a turn being used as test data with the remaining subsets being used as training data. At the end of the routine the performance metrics from across all folds are analysed (typically with the mean & standard deviation being reported) to provide a singular summary metric of general model performance.

A simple extension to this is repeated k-fold cross-validation, whereby the process of generating folds is repeated some number of times with the data being randomly shuffled before each new set of folds is generated. Thus, in repeated k-fold cross validation where k=4 and repeats=10, the data would be split into 4 folds 10 times (being randomly shuffled before each new set of folds) resulting in 40 models ultimately being trained and tested. Repeated k-fold cross validation has the advantage of further reducing the chance of overfitting by ensuring the makeup of every fold is not dependent upon the original order of the data.

2.3.2.5.2 Mean Absolute Error

A routine question in machine learning is how the quality of generated models should be assessed. Whilst r^2 is routinely used for regression models it is not suitable for use in ordinal regression which much of this thesis depends on (as the studied clinical scores which are used as dependent variables are almost always ordinal in nature). As such alternative measures are required for assessing regression model quality.

The Mean Absolute Error (MAE) belongs to a family of metrics that can be used to assess the quality of a predictive model (Chai and Draxler, 2014). MAE is calculated by summing the absolute difference between the predicted and actual values of a dependent variable for all predicted samples and dividing the result by the total number of samples as shown in the equation below.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Where n is the number of samples, y_i is the true value for instance i and x_i is the predicted value for instance i.

The Normalised MAE normalises the MAE with respect to the highest value of the dependent variable that is present in the dataset using the equation below (e.g., if the maximum value a dependent value *can* take is 28, but the maximum *observed* value is 18 the normalised MAE is calculated using 18, not 28). The normalised MAE allows comparisons between dependent variables with different scales. This is particularly useful here as it allows the comparison of model quality across clinical scores with differing scales (e.g., the UHDRS-TMS and TFC).

$$Normalised\ MAE = \frac{MAE}{\max(y)}$$

Where $\max(y)$ is the maximum score observed for dependent variable y.

For example, let two dependent values, A and B, with maximum observed scores of 5 and 10 respectively and two models use the same independent variable to predict each of A and B with an MAE of 1. The MAE of the models makes their performance appear equal; however, the normalised MAE will be 0.2 for dependent variable A and 0.1 for B, corresponding to an error of 20% and 10% respectively and showing that the dependent variable is better at predicting variable B than variable A.

It should be noted that an alternative commonly used method is the Root Mean Squared Error (RMSE) which can also be used to quantify the error in a prediction model. The MAE is preferred here to RMSE due to the increased weight RMSE gives to larger errors relative to smaller errors (e.g., it considers an error of 4 to be more than twice as bad as an error of 2 which does not translate well to predicting clinical values which are on a linear scale).

2.3.2.6 Step 5: Scatter plots and correlations between C3t scores (cross-correlation)

An issue with deriving scores from measured scores, as the C3t does extensively, is determining the added benefit of doing so. Whilst multiple C3t scores were discarded after step 1 and step 3, further refinement of the C3t was possible by investigating the cross-correlation between C3t scores.

C3t scores which are very highly correlated with each other are unlikely to provide any practical benefit to the test as they are, for all intents and purposes, the same score. This is also important to test as, in the event two or more scores are not highly correlated with each other (but are correlated with the same clinical measure), then high-quality multiple regression models may developed using them. As such, each C3t score correlated with the same clinical measure was extracted and Spearman's R correlations calculated between them. These are reported along with scatter plots showing the relationship between the relevant C3t score pairs.

2.3.2.7 Step 6: Effect Sizes

A key consideration when developing clinical assessments like the C3t is its performance over time. Ideally, clinical assessments should show altered performance in-line with disease progression. In the case of the C3t and HD, this translates to the C3t scores not changing over short time periods (e.g., days, weeks during which disease state is thought to be stable) but changing over longer time periods (e.g., months, years). Additionally, developed clinical assessment scores should ideally be similar to changes seen in any gold-standard assessment measures they have been found to be related to.

To test this in the C3t, effect sizes were used to assess the stability of the C3t scores over 1-month and 12-month periods (relative to the baseline visit). Effect sizes were also calculated for the UHDRS-TMS and CUHDRS between the baseline and 12-month visit, in order to assess whether changes in the C3t scores over the same time period were also seen in these clinical measures. The UHDRS-TMS and CUHDRS were not collected at the 1-month visit and so effect sizes for these measures between baseline and 1-month could not be studied. Effect sizes were not calculated for PIN_{HD} due to the very low sample size of the pre-manifest and prodromal cohort that had repeat visit data (n=3).

Traditionally effect sizes are used to notice change over time in longitudinal experimental data. The most popular technique for this arguably being Cohen's D (Cohen, 1988). However, Cohen's D is not applicable here as it assumes normality and many of the C3t scores were found to be non-normal in Step 2. As such, a non-parametric analogue of Cohen's D, Ω, was used instead (Wilcox, 2018).

2.3.2.7.1 A non-parametric analogue of Cohen's D

This section is based entirely on work by Wilcox (2018).

Cohen's D is calculated using the following equation.

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S}$$

Where $\bar{X}_i$ is the mean of variable X in group i and S is the pooled standard deviation of both groups.

In natural language, the (Cohen's D) effect size for a random variable, X, is the difference in mean value of X over two paired sets of observations divided by the pooled standard deviation of those two sets. This produces an effect size representative of how the two groups tend to differ (the mean difference) in terms of standard deviation units from across both sets of paired observations.

In the often cited source (Cohen, 1988) Cohen defines 'small, 'medium' and 'large' effect sizes as $d < 0.2$, $d < 0.5$ and $d < 0.8$ respectively. Critically, Cohen notes however that these effect sizes should be used as rough guides only and what constitutes a small/medium/large effect is entirely dependent on the phenomenon being studied.

Wilcox (2018) points out that Cohen's D provides a reasonable effect size estimate when normality and homoscedasticity are present. When these assumptions are not met however the validity of the measure starts to decrease.

Wilcox proposes an alternative calculation of effect size that does not rely on normality and homoscedasticity, Q, as shown below.

$$Q = F_0(\theta_D)$$

Where D is a distribution, $\theta_D$is the population median associated with D, and F_0 is the distribution of D when the null hypothesis that the median of the population, θ_D, is true.

In natural language, Q is defined as the extent to which the population mean θ_D represents a shift relative to F_0, away from 0 into a higher/lower quartile. Thus, in identical distributions $\theta_D = 0$ and $Q = 0.5$.

Wilcox continues further, defining Q relative to 0.5 as Ω calculated as follows.

$$\Omega = \frac{Q - 0.5}{0.5}$$

Ω will take on values from -1 to 1 and so $|\Omega|$ can be used to aid interpretation when the direction of the effect is not important.

Finally, Wilcox defines small, medium, and large descriptor analogues of Cohen's D for values of $|\Omega|$ as $|\Omega| = 0.1, 0.3$ and 0.4 respectively. Ω is reported in this study in place of Cohen's D when estimating effect sizes of variables from baseline to 1-month and baseline to 12-months.

2.4 Results

2.4.1 Participants

One-hundred and five gene-positive participants were recruited at the baseline visit across all studies and sites of which thirty-three have follow-up visits. Some participants had missing C3t scores. Demographics at the baseline visit are shown in Table 16, the sample size of each C3t score is shown in Table 17.

Table 16: Participant demographics per TFC stage and over the whole cohort. Other than n= and % Female, given values are the mean (± standard deviation).

TFC Stage Group	n=	% Female	Age	CUHDRS	PIN_{HD}	UHDRS-TMS	Whole-body Chorea	Upper-body Chorea
Pre-manifest	5	20	43.0 (±9.8)	17.3 (±2.0)	-0.2 (±0.9)	0.4 (±0.5)	0.0 (±0.0)	0.0 (±0.0)
Prodromal	11	36.4	47.5 (±12.6)	14.1 (±2.7)	1.0 (±1.4)	10.8 (±6.1)	2.2 (±1.5)	0.8 (±1.0)
TFC Stage 1	39	25.6	54.4 (±11.3)	12.0 (±2.1)	N/A	25.2 (±12.2)	9.1 (±4.2)	5.2 (±2.5)
TFC Stage 2	43	51.2	52.7 (±12.0)	7.6 (±2.2)	N/A	39.4 (±12.7)	10.7 (±4.6)	6.1 (±3.0)
TFC Stage 3	7	28.6	46.1 (±8.9)	4.6 (±2.5)	N/A	43.7 (±22.2)	9.1 (±7.7)	5.0 (±4.6)
Whole cohort	105 (n=16, PIN_{HD})	37.1	51.9 (±11.8)	10.2 (±3.8)	0.61 (±1.4)	29.6 (±17.0)	8.6 (±5.3)	4.9 (±3.3)

Table 17: Sample size of participants with present data for each C3t score used in this study at baseline.

C3t Score	Sample Size
BTT Time Taken	105
BTT Dropped Tokens	87
BTT Rule Errors	105
BTT Transfer Errors	105
BTT Accuracy Score	105
BTT Total Task Score	87
CTT Time Taken	105
CTT Dropped Tokens	87
CTT Rule Errors	87
CTT Transfer Errors	87
CTT Accuracy Score	87
CTT Total Task Score	87
CTT Time Performance Cost	105
CTT Task Score Performance Cost	87

2.4.2 Histograms

Table 18 shows the C3t scores removed & retained and the histogram plots for the BTT and CTT scores are shown Figure 2 and Figure 3 respectively.

In both tasks, the dropped token, rule error and transfer error scores are highly invariant, with almost all participants not dropping a single token or making a rule/transfer error. As such these scores are unlikely, by themselves, to contain useful information about the population and so were removed from subsequent analysis steps. The BTT and CTT accuracy scores are also highly invariant and are derived solely from the transfer and rule errors. As such both the BTT and CTT accuracy scores were removed from subsequent analysis steps as well.

The BTT and CTT total task scores are derived from the time taken measures, the number of dropped tokens and the accuracy score. All but one of these component scores (time taken) were removed, however the total task score was retained in both tasks to see if the very small variance the errors it includes display proved useful when combined with the time taken measures.

The time taken and performance cost measures for both tasks were highly varied and so were retained.

Table 18: Retained and removed C3t measures following visual inspection of distributions.

C3t Score	Status
BTT Time Taken	Retained
BTT Total Task Score	
CTT Time Taken	
CTT Total Task Score	
CTT Time Performance Cost	
CTT Task Score Performance Cost	
BTT Dropped Tokens	Removed
BTT Rule Errors	
BTT Transfer Errors	
BTT Accuracy Score	
CTT Dropped Tokens	
CTT Rule Errors	
CTT Transfer Errors	
CTT Accuracy Score	

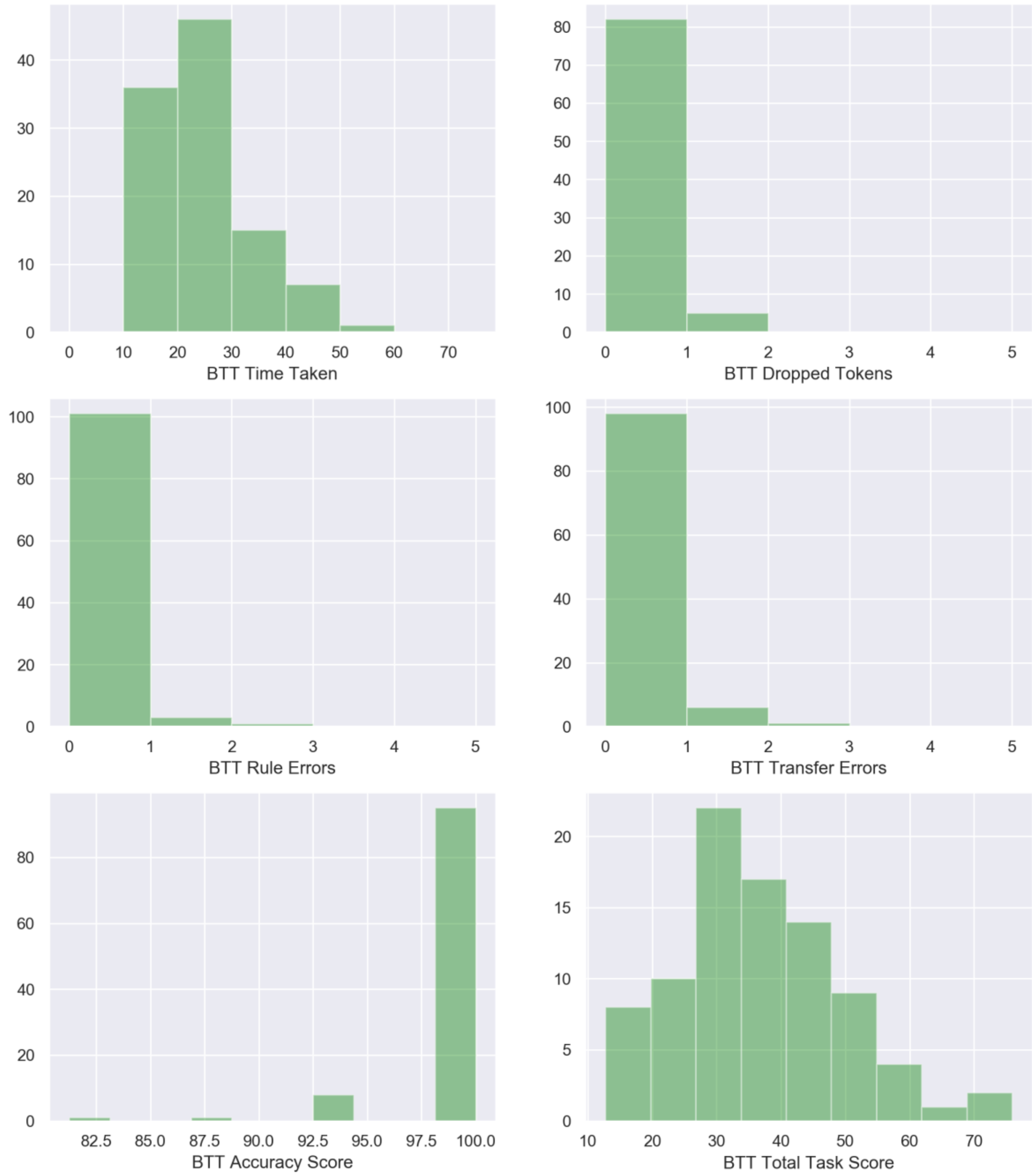


Figure 2: The distribution of each BTT score over the whole cohort.

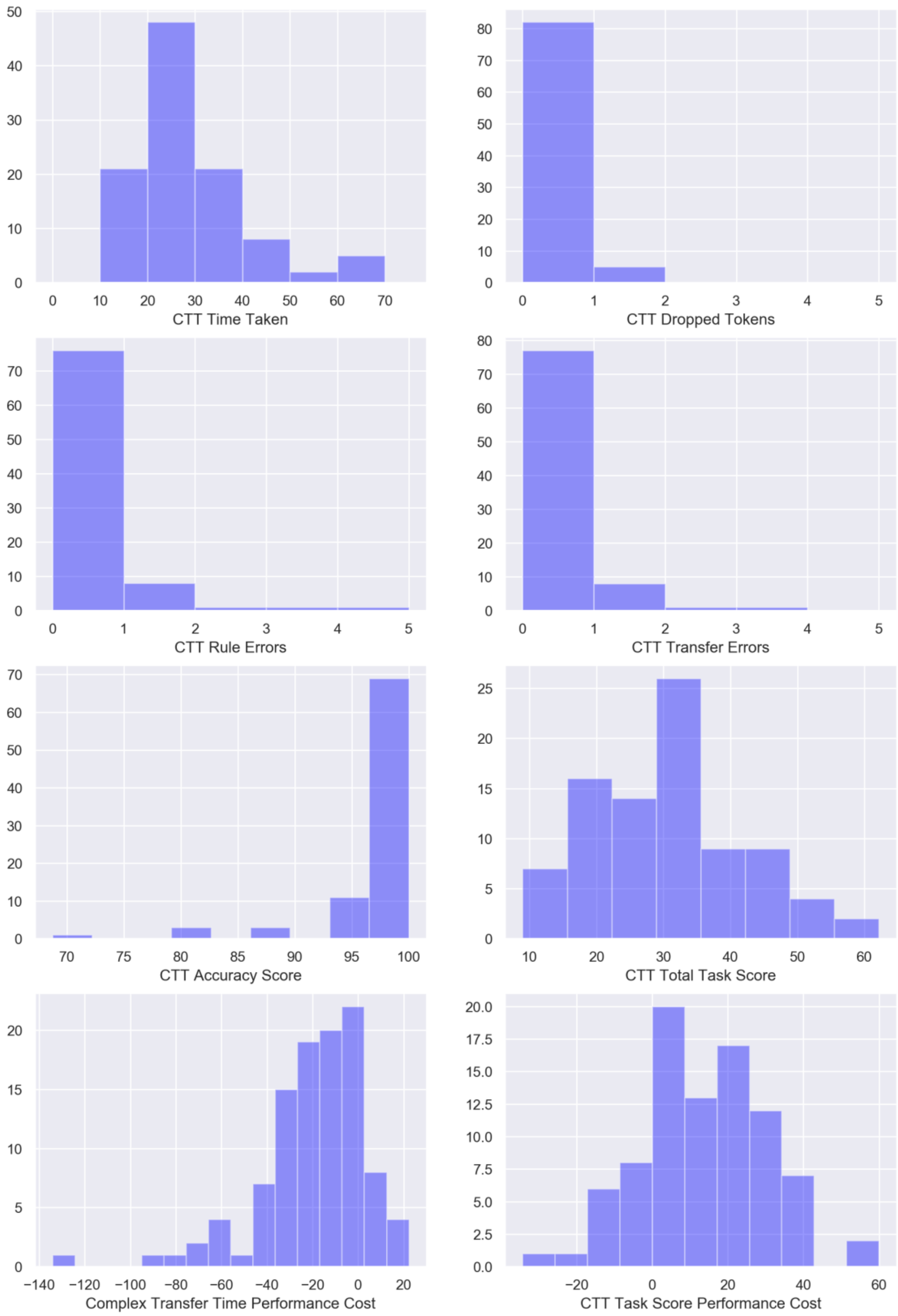


Figure 3: The distribution of the CTT scores over the whole cohort.

2.4.3 Scatter Plots

Monotonic relationships were detected for the BTT and CTT time taken scores and total task scores with the CUHDRS, UHDRS-TMS and PIN_{HD}.

The CTT time performance cost and task score performance cost scores had no relationship with any clinical score. No C3t score appeared to have a monotonic relationship with whole-body chorea or upper-body chorea.

Table 19 shows which pairs of scores appear to have monotonic relationships. Figures Figure 4, Figure 5, Figure 6, Figure 7, and Figure 8 show the scatter plots for the C3t scores and the CUHDRS, UHDRS-TMS and PIN_{HD}, whole-body chorea and upper-body chorea respectively.

The two CTT performance cost scores were removed at this point from consideration, as they were seen to be clearly unrelated related to any clinical measure (see figures 3-7).

Table 19: The presence of a monotonic relationship between C3t scores and clinical scores is indicated with an 'x', its absence is indicated with a dash (-).

C3t Scores	Clinical Scores				
	CUHDRS	UHDRS-TMS	PIN_{HD}	Whole-body chorea	Upper-body chorea
BTT Time Taken	x	x	x	-	-
BTT Total Task Score	x	x	x	-	-
CTT Time Taken	x	x	x	-	-
CTT Total Task Score	x	x	x	-	-
CTT Time Performance Cost	-	-	-	-	-
CTT Task Score Performance Cost	-	-	-	-	-

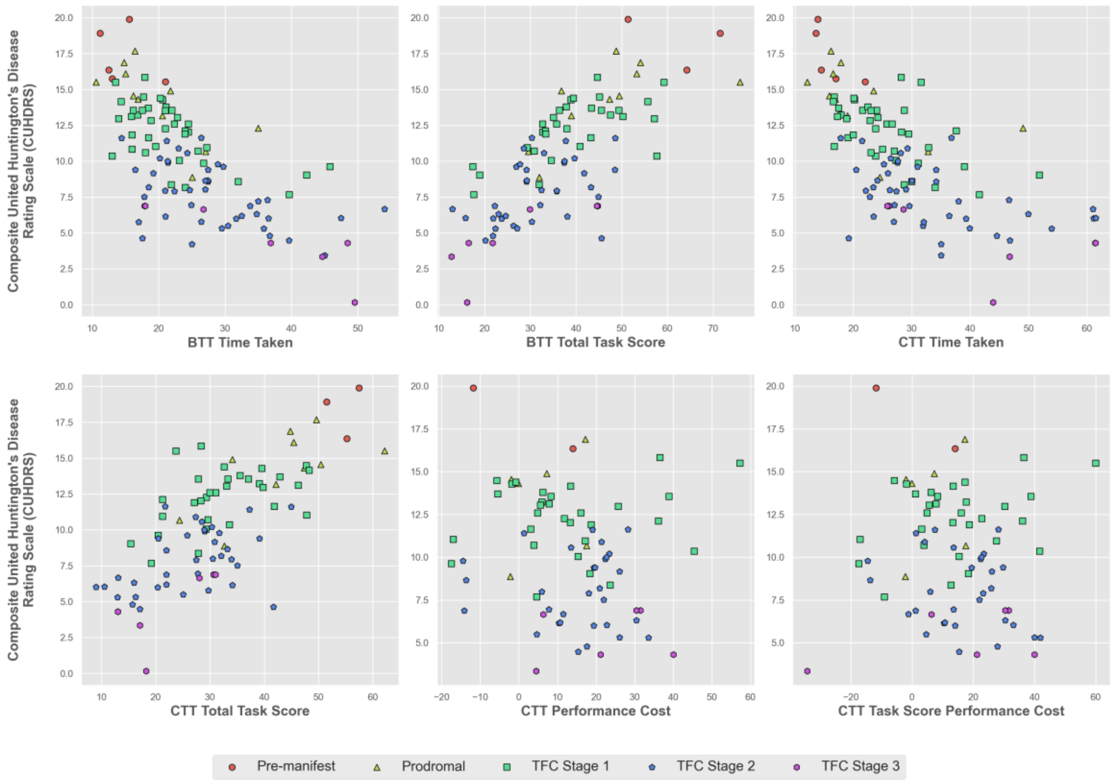


Figure 4: Scatter plots of each retained C3t score with the CUHDRS

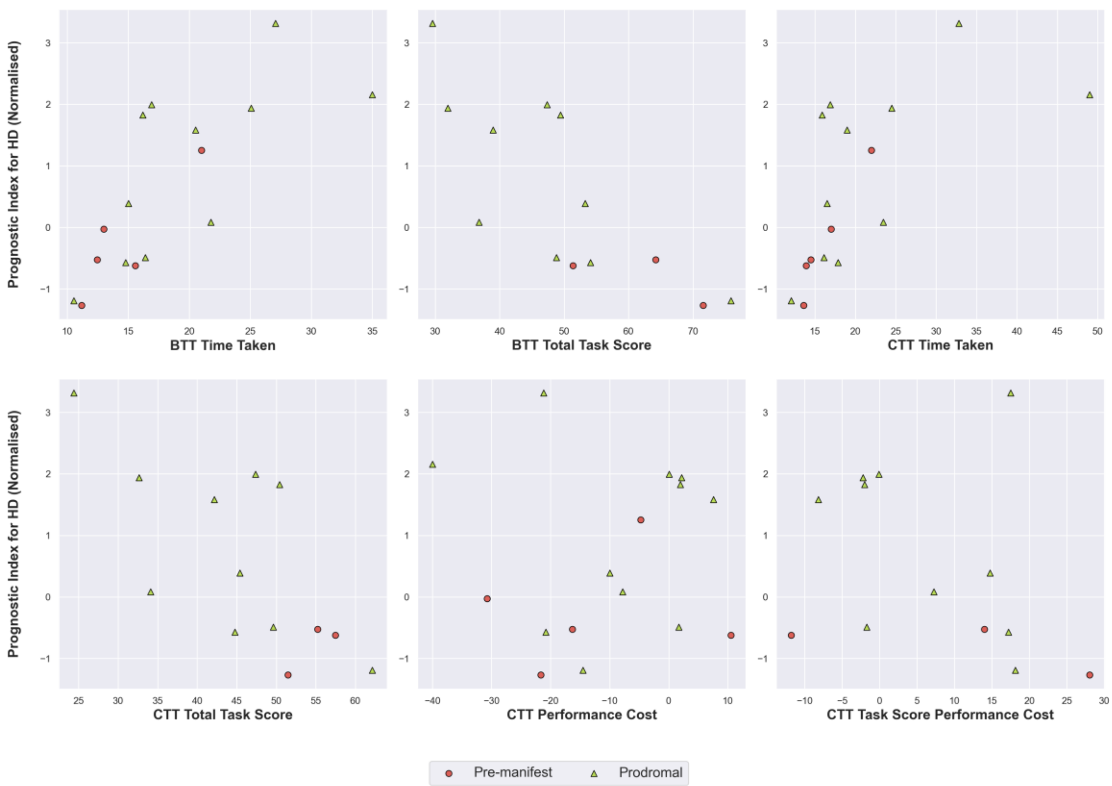


Figure 5: Scatter plots of each retained C3t score with PIN_{HD}

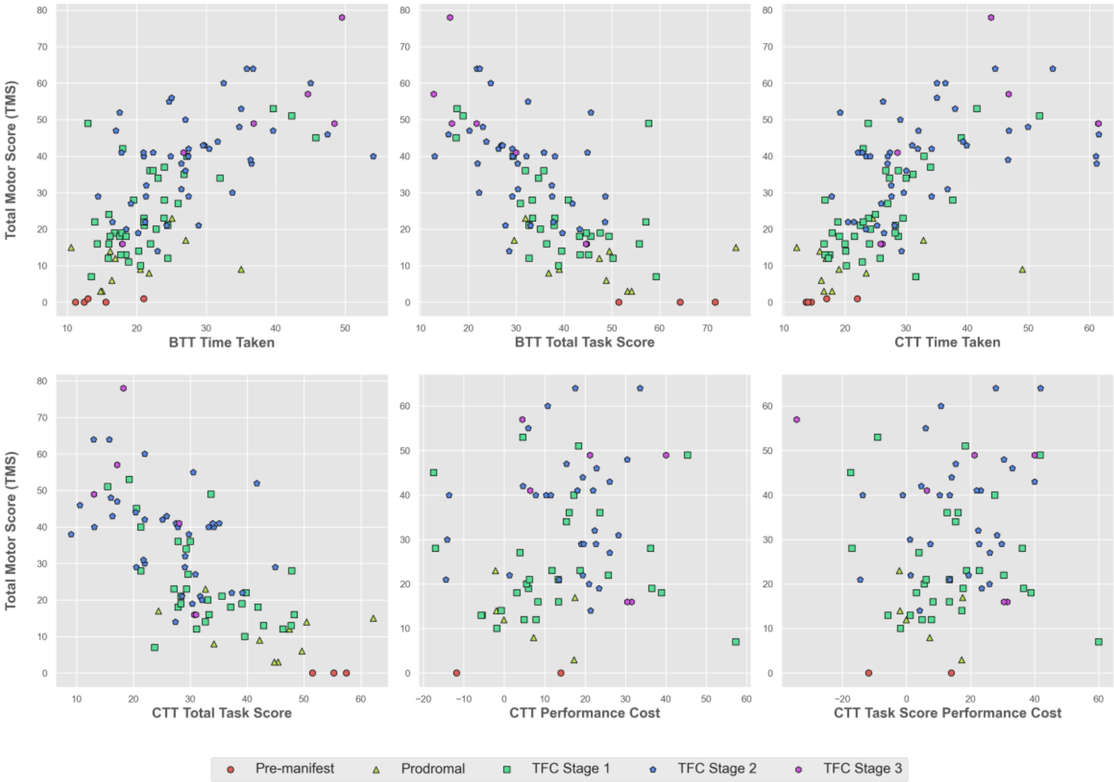


Figure 6: Scatter plots of each retained C3t score with the UHDRS-TMS

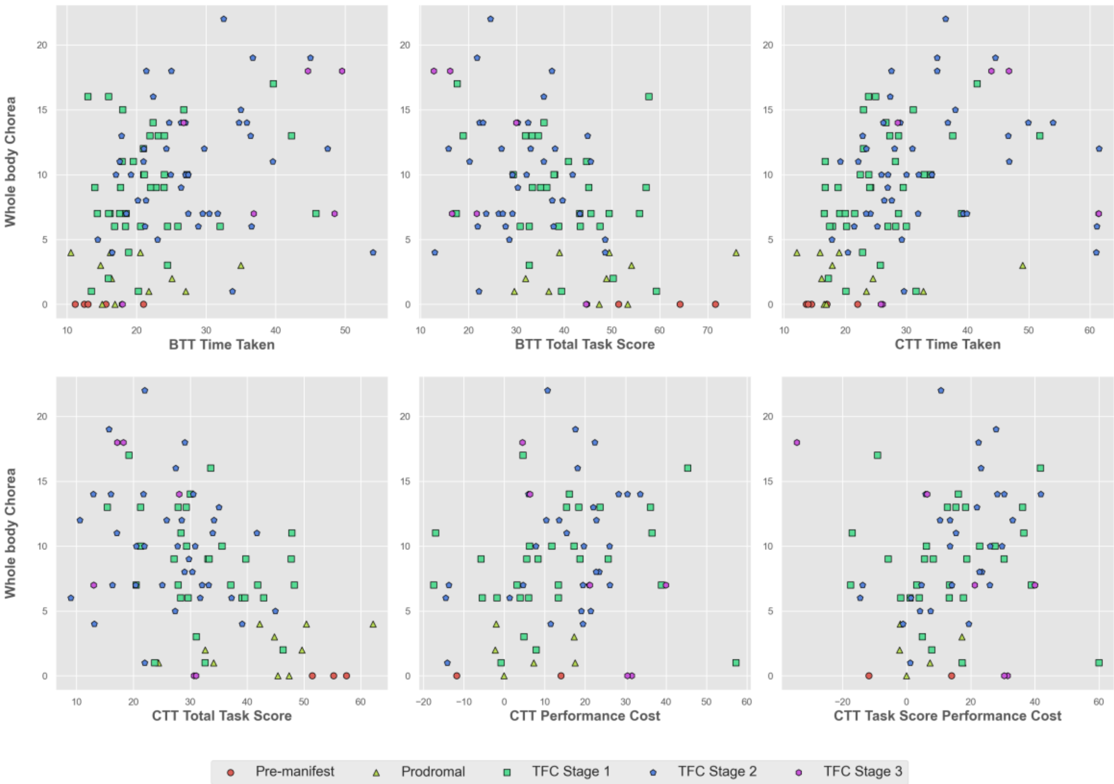


Figure 7: Scatter plots of each retained C3t score with whole-body chorea

Figure 8: Scatter plots of each retained C3t score with upper-body chorea

2.4.4 Correlation & Regression

As no monotonic relationships were observed between any C3t score and whole-body chorea, or upper-body chorea correlation and regression analyses were not run for these clinical measures and hence were not investigated any further. Correlation and regression analysis was run between the retained C3t scores (BTT time taken, CTT time taken, BTT total task score, and CTT total task score) and the CUHDRS and UHDRS-TMS. Correlation analysis only was for the C3t scores and PIN_{HD}.

Strong correlations were found between the time scores and total task scores for both C3t tasks and the CUHDRS, UHDRS-TMS and PIN_{HD}. All Holm-Bonferroni corrections were passed for all variables, indicating a high chance of valid statistical significance. The performance of the regression models was good, with normalised MAE scores ranging from 11% to 13%.

As is shown in the correlation and regression results reported in Table 20, no one C3t score greatly outperformed the others. The sites and test versions were found to not impact model quality, with coefficients of the both significantly lower than that of the C3t score variables for all models & clinical scores. Full coefficient results can be found in appendix 6.1, tables Table 45, Table 46, Table 47, and Table 48.

Table 20: Correlation and regression results for the retained C3t scores and UHDRS measures. The mean (± standard deviation) is reported for the Mean Absolute Error (MAE). The Normalised MAE (N-MAE) was calculated from the mean MAE score. n=16 for all PIN_{HD} analysis (pre-manifest & prodromal participants only); n=105 for BTT Time Taken & CTT Time Taken; n=87 for BTT Total Task Score and CTT Total Task Score.

C3t Scores	CUHDRS			UHDRS-TMS			PIN_{HD}		
	r	MAE (± std)	N-MAE	r	MAE (± std)	N-MAE	r	MAE (± std)	N-MAE
BTT Time Taken	-0.69***	2.23 (±0.33)	11.0	0.67***	9.88 (±1.43)	13.0	0.83***	N/A	N/A
BTT Total Task Score	0.73***	2.2 (±0.3)	11.1	-0.72***	9.4 (±1.8)	12.1	-0.82**	N/A	N/A
CTT Time Taken	-0.7***	2.11 (±0.3)	11.0	0.69***	9.4 (±1.18)	12.0	0.76**	N/A	N/A
CTT Total Task Score	0.69***	2.2 (±0.3)	11.0	-0.7***	9.7 (±1.9)	12.4	-0.7**	N/A	N/A

2.4.5 Scatter plots and correlations between C3t scores

Both the BTT and CTT time scores were very highly related to their corresponding total task scores (r=-0.99). This is expected given that the total task scores are derived from the time scores and the other scores that are used to calculate the total task scores (dropped tokens, transfer errors and rule errors) were found to be invariant in section 2.4.2.

The BTT and CTT time scores are also strongly related to each other (r=0.86). This seems however to apply less as the scores increase, creating the 'funnel' shape that can be seen in Figure 9 (top left plot). As the BTT and CTT time scores are highly correlated and are each extremely correlated with their respective total task scores, it is unsurprising that similarly the BTT and CTT total task scores were found to be similarly strongly correlated with each other (r=0.86).

Figure 9 shows scatter plots and associated Spearman's R correlation statistic for each retained C3t score pair.

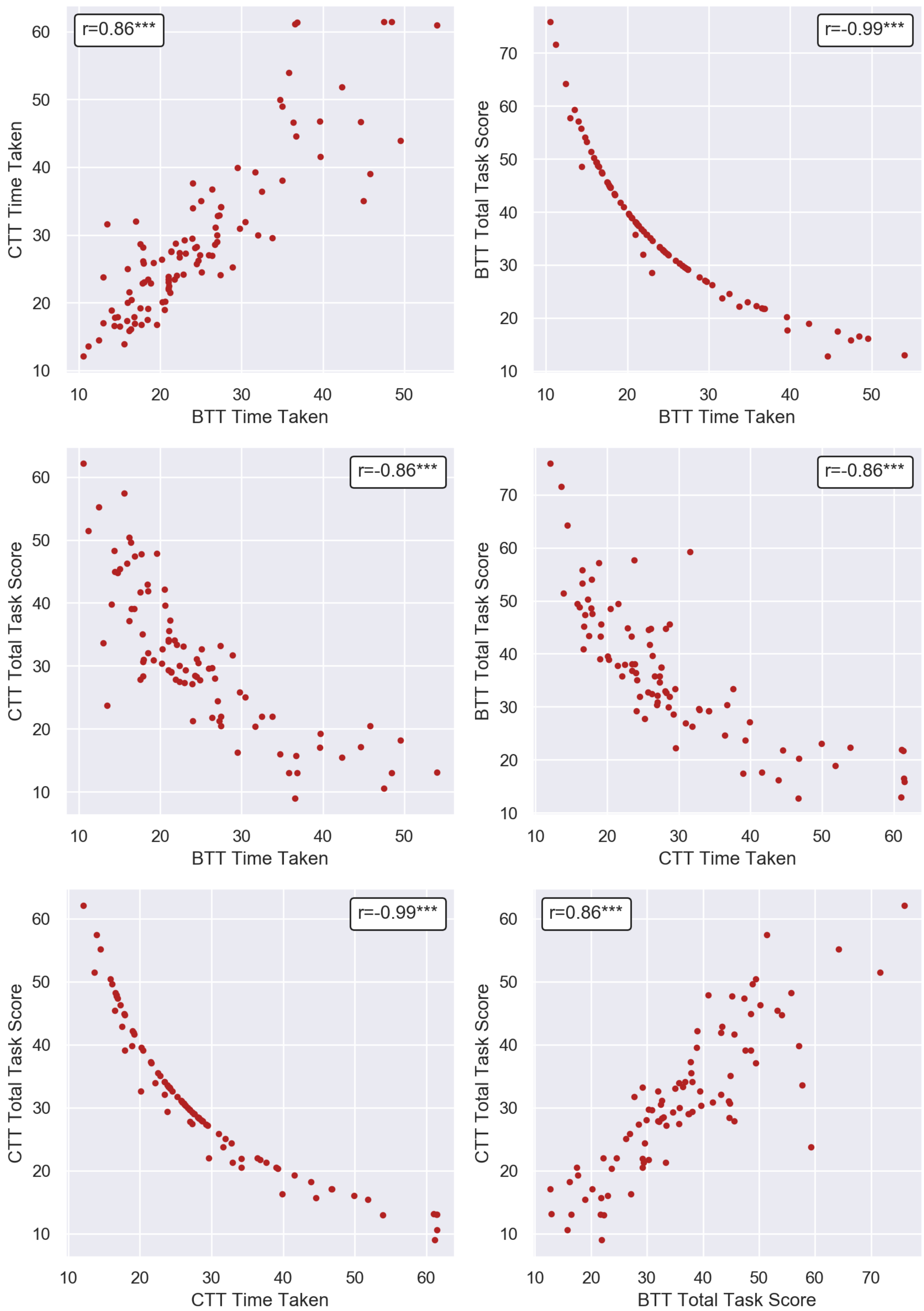


*Figure 9: Relationship between each of the C3t scores. The Spearman's R value for each is shown. P<0.001 in all cases (indicated by ***).*

2.4.6 Effect Sizes

Changes in the BTT and CTT time scores from baseline to 1-month were small (|Ω|<0.3). Similarly, small changes in both the BTT and CTT time scores and the CUHDRS and UHDRS-TMS were seen from baseline to 12-months (|Ω|<0.3). Effect sizes for each measure are shown in Table 21.

Effect sizes were only calculated for the C3t time scores as at this point in the analysis all other analysed scores could be safely discarded. Due to some missing data, the sample size for the CUHDRS was reduced from n=33 to n=22.

Table 21: Non-parametric effect size, Ω, for the BTT Time Taken in Seconds, CTT Time Taken in Seconds, CUHDRS and UHDRS-TMS

Score	Time Period	
	Baseline to 1-month Ω (n=)	Baseline to 12-months Ω (n=)
BTT Time Taken in Seconds	0.008 (n=33)	-0.06 (n=33)
CTT Time Taken in Seconds	0.095 (n=33)	0.1 (n=33)
CUDHRS*	N/A	0.062 (n=22)
UHDRS-TMS	N/A	-0.073 (n=33)

2.4.7 C3t score removal

Throughout the process of this study numerous BTT and CTT scores were removed from further analysis. With the analysis concluded it is possible to make final recommendations as to which should be retained, and which should be removed.

The rule errors, transfer errors, and dropped token scores for both the BTT and CTT are recommended for removal. This is because these six scores were all found to be highly invariant via visual analysis in the studied cohort with almost all participants making no errors. As the accuracy scores for the BTT and CTT are derived solely from the rule errors and transfer errors the accuracy scores are also recommended for removal.

The time performance cost scores for the BTT and CTT are recommended for removal as neither was found to have a monotonic relationship with any clinical measure.

The time taken scores and total task scores for the BTT and CTT were all found to vary throughout the cohort, were highly correlated with various clinical measures and produced strong predictive models. However, cross-correlation between the time scores and their respective total task scores was extremely high. Due to the added complexity required to calculate the total task scores, and the fact that neither showed clear benefits over the time taken scores in terms of correlation strength or predictive model performance, the total task scores are recommended for removal.

The time taken scores are recommended for retention, although it is unlikely both are needed as they show similar relationships with the clinical measures and are themselves highly correlated. More information however is required before making a recommendation as to which should be retained.

Table 22 summarises the recommendation for each score along with a brief rationale as to why.

Table 22: Recommendations for the retention/removal of each C3t score with the reason for the recommendation.

<table>
<tr><th>C3t Score</th><th>Recommendation</th><th>Reason</th></tr>
<tr><td>BTT Time Taken</td><td rowspan="2">Retain</td><td rowspan="2">Sensitive to clinical measures</td></tr>
<tr><td>CTT Time Taken</td></tr>
<tr><td>BTT Rule Errors</td><td rowspan="14">Remove</td><td rowspan="6">Invariant</td></tr>
<tr><td>BTT Transfer Errors</td></tr>
<tr><td>BTT Dropped Tokens</td></tr>
<tr><td>CTT Rule Errors</td></tr>
<tr><td>CTT Transfer Errors</td></tr>
<tr><td>CTT Dropped Tokens</td></tr>
<tr><td>BTT Accuracy Score</td><td rowspan="2">Invariant and derived solely from invariant measures</td></tr>
<tr><td>CTT Accuracy Score</td></tr>
<tr><td>CTT Time Performance Cost</td><td rowspan="2">No monotonic relationship with clinical measures</td></tr>
<tr><td>CTT Task Score Performance Cost</td></tr>
<tr><td>BTT Total Task Score</td><td rowspan="2">Very high correlation with respective time scores, no added value compared to using time scores alone, added complexity to calculate</td></tr>
<tr><td>CTT Total Task Score</td></tr>
</table>

2.5 Discussion

The results presented here show that a subset of the of the studied C3t scores are strongly correlated with the CUHDRS, UHDRS-TMS, and PIN_{HD}.

Notably, strong correlations between C3t time scores and the UHDRS-TMS were observed, but such correlations were not seen with whole-body chorea or upper-body chorea (components of UHDRS-UHDRS-TMS). Unfortunately, is shown in Figure 7 and Figure 8 no C3t score had an observable monotonic relationship with whole-body or upper-body chorea. As such, correlation, and regression analysis were not performed for either measure of chorea as the results from such analysis would be invalid.

Given the known link between the C3t scores and the UHDRS-TMS (Clinch *et al.*, 2018), the complete lack of an observable relationship with chorea was surprising. Whilst a reduction in correlation strength was expected, the total absence of any observable relationship was not. It is thus necessary to consider what aspect of HD symptoms, particularly motor symptoms, the C3t scores are sensitive to. As only the BTT and CTT time taken scores, which measure the length of time each task takes, were found to be related to the clinical measures only these scores will be discussed. Please note that whilst the total task scores were related to the clinical measures, they were so strongly related to their respective time scores that they are effectively the same measure. As such there is no reason to retain nor discuss them in this context.

The impact HD will have on the time taken for participants to complete the BTT and CTT is simple – as symptom severity increases so too will the time participants take to complete the tasks. As has already been stated, this relationship has been previously reported & observed for measures of cognitive deficits, functional capacity and, as has been repeated here, motor symptoms via the UHDRS-TMS (Clinch *et al.*, 2018).

Cognitive symptoms will logically lead to indirect increases in the time it takes to complete C3t transfer tasks. For example, a reduction in attention span might cause participants to not pay full attention to performing the task or psychomotor slowing may inhibit participants ability to move swiftly through the different stages of the task. Motor symptoms meanwhile will increase the time scores more directly. For example, hyperkinetic disturbances will prevent the dexterity-based aspects of the tasks from being completed smoothly whilst hypokinetic disturbances will prevent the tasks from being completed quickly. As the reduction in functional capacity is an emergent property of the other symptom domains seen in HD, and those component symptom domains are related to C3t task times, a reduction in functional capacity should also correspond to increased task times, as has been observed. Whilst the C3t being related to multiple symptom domains makes sense, it does not explain why the C3t time scores appear related to the UHDRS-TMS but not the chorea sub-scores. A possible explanation for this, however, lies in the makeup of C3t time scores, the UHDRS-TMS, and the nature of motor symptoms in HD.

HD motor symptoms evolve over time, with hyperkinetic disturbances (like chorea) existing in the earlier stages of the disease which eventually give way to hypokinetic disturbances later on (McColgan and Tabrizi, 2018). The UHDRS-TMS accounts for this evolution by summing the result of multiple individual motor assessments, only one of which is chorea (Kieburtz *et al.*, 1996). The individual motor symptoms which can present during HD could be logically expected to impact the time each task takes in different ways. For example, chorea could make it harder to pass tokens between hands quickly, oculomotor dysfunction could make identifying the next token and placing it into the slot harder, and bradykinesia could simply slow down the rate at which participants can move. The difficulty is these symptoms may present in a patient at the same time, and the C3t time scores have no mechanism by which they can notice which symptom(s) are causing the increase in task time (Roos, 2014). As an example, one patient for example might have low levels of chorea but high levels of bradykinesia, and another the reverse. Both patients would be expected to take longer to perform the C3t, but the C3t time scores do not contain enough information to differentiate between the two patients.

In general, as motor symptoms worsen one would expect C3t performance to be progressively reduced, and therefore one would expect a link between C3t performance and general motor function/dysfunction. This is exactly what is seen here – the C3t appears sensitive to the UHDRS-TMS, a summary score of general motor function, but insensitive to specific symptoms, in this case chorea. Given that we also know that the C3t time scores are impacted by cognitive deficits as well as motor deficits, the lack of a direct relationship between C3t time scores and individual motor symptoms makes all the more sense.

This property of the C3t makes the instrumentation of the test all the more worthwhile. However, it does mean that a different baseline for the success of instrumentation will have to be selected, as will be discussed in chapter 4.

The proposition that the C3t is strongly related to general HD disease state rather than specific symptoms, is further evidenced by the second finding of this study, namely that the C3t is strongly correlated with both the CUHDRS and PIN_{HD} and can estimate to a high degree of accuracy the CUHDRS. In practical terms, this shows the C3t is both affected by general disease state (the CUHDRS) and is related to *estimated* progression levels (PIN_{HD}).

This quality of the C3t is potentially useful for clinical trials, although the test itself is currently a clinical research tool that should be considered in development. Clinical trials require the regular large-scale collection of measures known to be sensitive to progression. The CUHDRS is known to have superior sensitivity (relative to its component assessments) to disease progression. A limitation of the CUHDRS however is that it requires four individual assessments from the UHDRS in order to be calculated, meaning that patients must be assessed by those trained to administer the UHDRS. Whist this is not a problem in infrequent, small-scale studies it makes highly regular (e.g., daily, weekly) data collection in large scales impractical.

Unlike the UHDRS however the C3t requires minimal training and could conceivably be conducted by the patient's primary carers or family members at home. Whilst this would not replace the need for regular gold-standard clinical evaluation it would allow the CUHDRS to at least be approximated at a regularity and scale otherwise impractical.

The C3t was found to be highly correlated with PIN_{HD} however due to the low sample size available it was not felt appropriate to run regression models for PIN_{HD}. Future work may wish to collect additional pre-manifest and prodromal data and assess whether the C3t can be used to estimate PIN_{HD}. If the C3t can be used to accurately estimate PIN_{HD} in the same manner that it has been shown here to be able to estimate the CUHDRS, then the C3t could be used to regularly estimate PIN_{HD} in the home. This could be used to further show the validity of PIN_{HD} by studying its progression over time. Unlike the CUHDRS however PIN_{HD} is a measure of projected progression rather than a measure of disease state. Whilst its regular approximation may give valuable insights into how PIN_{HD} evolves over time its regular approximation using the C3t is arguably less useful than regularly approximating the CUHDRS would be. Regardless, more work is needed to determine whether the C3t can be used to estimate PIN_{HD} as it can the CUHDRS.

A dataset containing truly regularly approximated measures of general HD disease state would be invaluable to research. The potential to observe a slow but steady increase in symptom severity over the course a year could provide valuable insights into how HD evolves over time.

Whilst regularly approximating the CUHDRS and PIN_{HD} would be highly valuable to HD research, there are however two questions that would need to first be addressed. Firstly, whether there is a training effect from repeatedly taking the C3t, as if a training effect does exist it would have to be taken to account for it when estimating measures, and secondly whether the C3t is sensitive to change over time.

These questions were planned to be answered by the analysis presented here, however the C3t appears stable over short time-periods but whether it is truly anchored to clinical measures is still unknown.

The effect size, Ω, for the C3t time scores fell into the category of 'small' ($|\Omega|<0.3$) from baseline to 1-month and baseline to 12-months. The effect size for the CUHDRS and UHDRS-TMS was also small ($\Omega<0.3$) from baseline to 12-months. Over a period of 1-month, HD is not expected to noticeably progress (this being the reason the UHDRS was not conducted and so the CUDHRS and UHDRS-TMS were not available at 1-month). As such the lack of progression in the C3t would be expected if it is anchored to these measures.

Over a period of 12-months the CUHDRS and UHDRS-TMS are known to typically progress to some degree (Tabrizi *et al.*, 2011; Schobel *et al.*, 2017). In our cohort however such progression was not observed in any clinical measure analysed or in the C3t scores. The lack of progression in the observed cohort for the CUHDRS and UHDRS-TMS should be considered as atypical, although it could be due to the small sample size available to this study.

We can conclude, at most, that in our small cohort the C3t appears to be stable over short time periods when progression would not be expected to occur (baseline to 1-month) and did not progress over 12-months when clinical measures did not progress either. Unfortunately, we cannot conclude that the C3t scores are anchored to changes in the CUHDRS or UHDRS-TMS This is because to confirm the C3t scores are anchored to changes in those measures we must be able to observe the C3t scores changing over time with them, which we have not. The small sample size adds to the inconclusiveness of these results.

Parallel to this is the issue of practice effects, the phenomena of changes in test performance (typically improved performance) due to participants having prior practice with or exposure to the test (McCabe *et al.*, 2011). In the context of the C3t, any practice effects would presumably result in a reduction in the time it takes participants to complete the tasks. Practice effects could conceivably exist both within individual C3t tests (due to the mechanically very similar C3t transfer tasks), and between C3t tests if performed multiple times. If practice effects were found to exist in the C3t this would naturally have implications for how it could be used as a research tool, particularly for longitudinal studies.

There is evidence in the literature that suggests the C3t may be subject to a practice effect. The nine hole peg test (a common clinical assessment where participants transfer nine pegs into a series of slots as fast as possible using either their dominant or non-dominant hand (repeated for each hand)) was found to have a significant practice effect in both a control population and a population of participants with multiple sclerosis (Mathiowetz *et al.*, 1985; Solari *et al.*, 2005). This has been controlled for in some studies by having participants purposely repeat the test numerous times to account for the effect, however it has also been argued that this may be needless given how often multiple sclerosis participants are exposed to the test (Feys *et al.*, 2017).

Overall, establishing whether the C3t is anchored to clinical progression and whether it is subject to practice effects are both crucial for determining the worthiness of the C3t as a clinical research tool, particularly regarding the concept of regularly approximating other measurements. If the C3t is not anchored to the progression of those measures, then their regular approximation would be effectively meaningless. Similarly, if practice effects were found either between C3t tests or within individual C3t tests this would need to be addressed in order for true test performance results to not be obscured. As such, it is crucial that for the continued development of the C3t a proper study into its longitudinal properties be conducted and these questions explored. The lack of exploring practice effects in the C3t is a recognised limitation of this study, and as such is one of the recommendations made for future work.

Whilst there are still questions to be answered regarding the C3t the analysis presented here however does show that numerous C3t scores can be removed without impacting its efficacy as a clinical measurement. The various analysis steps and results of this study suggest together that the C3t can be substantially simplified by removing all measured and derived scores (from the BTT and CTT) other than the time taken measures. Doing so is desirable as it makes the C3t simpler to administer which would be particularly helpful if the concept of having regular in-home collection by non-clinicians performed is pursued.

Table 22 shows the full list of C3t scores analysed during this study, whether they were retained or removed and if removed why.

Rule errors, transfer errors, and dropped tokens can be removed from both tasks as the histograms in Figure 2 and Figure 3 show that across the whole cohort the vast majority of participants had perfect scores for each of these measures. As the accuracy score of both tasks are computed from their respective transfer and rule errors, and these are by-large invariant, the accuracy scores may also be removed.

The total task scores (calculated from accuracy, dropped tokens, and time taken) and performance cost scores (time & total task score difference between the BTT and CTT) were retained in case the small amount of variation in the errors did prove useful when combined with time.

Neither the time performance cost scores nor the total task score performance cost scores showed a monotonic relationship with any clinical score. As such we can conclude that the difference between performance in the BTT and CTT has no relationship with symptom severity. The lack of any relationship here is interesting as it shows that how participants perform on the CTT relative to the BTT is seemingly not determined by their symptoms. It is possible that the difference in task difficulty between the BTT and CTT is insufficient for the tasks to be impacted differently by different symptom severity levels. Either the BTT needs to be made easier or the CTT needs to be made harder. Unfortunately, the DTT, which is supposed to be harder than both the BTT and CTT was not assessed in this study due to sample size limitations. Such analysis may provide the answer as to how the test should be updated in the future such that performance gaps between the tests can be seen. Ultimately more analysis would need to be conducted to understand this further, however as it stands both types of performance cost scores can likely be removed without negatively impacting the C3t.

The time scores and task scores all showed strong, significant correlations with the CUHDRS, UHDRS-TMS and PIN_{HD} and could estimate the CUHDRS and UHDRS-TMS with a low level of error. The difference in correlation strength and predictive accuracy between the C3t scores was minimal. Further analysis showed very strong correlations between all four of these C3t scores.

The correlation between a time score and its associated total score is unsurprising as the latter is derived from the former. As shown by the histograms, the other variables of the total score are effectively invariant making the total task score a transformation of the time score. As such, taking into consideration the added complexity of recording the additional scores needed to compute the total task scores, the total task scores should be removed from the BTT and CTT.

The correlation between the BTT and CTT time scores deserves further investigation. As was observed, the difference in performance (in terms of time) between these tasks was unrelated to any clinical measure. The BTT-CTT time score scatter plot in Figure 9 (top left) shows the relationship between them seems to degrade, fanning out as the scores take on more extreme values. Whilst this may be due to random chance this 'fan' effect should be investigated as there may be an as-yet uncovered reason for this effect. One potential explanation is the added cognitive component of the CTT, and/or the lack of such a component in the BTT makes the tasks respectively more sensitive to some aspect of HD not investigated in this study.

As mentioned earlier, the C3t cannot be used to distinguish between symptoms domains, being sensitive to all of them simultaneously. However, the studied cohort consists of primarily early-stage manifest participants. It is possible that at the more extreme ends of the symptom spectrum this may stop being the case with the tasks sensitivity to different symptoms either increasing or decreasing. If the tasks were found to be sensitive to more severe symptoms this would be useful as there is a recognised need for assessment development sensitive to late-stage HD (Youssov *et al.*, 2013). This is also the case with the three baseline tasks, the BVT, CVT, and BAT, which were also excluded from this study. The rationale was excluding the baseline tasks was slightly different, being due to the lack of a movement component it is unlikely they would be impacted by movement disorders. Whilst this rationale seems reasonable, this does mean again that if it is felt that the baseline tasks should be looked at then this analysis would need to be conducted again.

The final question is which, given the strong correlation between the BTT and CTT time scores and their similar performance for predicting clinical measures, which should be retained.

Ultimately which score should be retained will depends on the use case. If the C3t were to be put into homes for regular collection the BTT should be preferred due to its simplicity. On the other hand, if there was a need to specifically assess the impact of cognition on motor function the CTT should be preferred.

In general, however, for future research it would be wise to retain both scores until a firm rationale for which should be removed (if either) is found. It is likely that either the BTT or CTT needs to be made either easier or harder respectively, as the difference in performance between the two was unrelated to any clinical measure. Once one of the tasks has been updated accordingly this result may change. As such, for research whilst the other BTT and CTT scores can be safely removed, we recommend both tasks and time scores be retained for now.

Finally, it was found that neither which site the C3t was conducted in, nor which test version was used, influenced regression model quality for the CUHDRS and UHDRS-TMS relative to the C3t scores. Whilst not a core question investigated by this study, the reliability of the C3t between different study sites and test versions is important to consider as the C3t matures. As will be discussed in chapter 3, modern research requirements more and more often require data to be pooled from across multiple study sites and from previous studies in order to increase available sample sizes and so enable robust statistical analysis and modern machine learning techniques. This is particularly the case in research involving rare disease such as HD, where any given study site may have limited population sizes available for recruitment. Similarly, by including data from previous projects sample sizes can similarly be increased. As such, clinical assessments such as the C3t need to be standardised such that variations caused by data being collected by different projects, researchers, teams, and sites are minimal to non-existent. This study provides the first quantitative evidence that the effect of pooling C3t data from multiple sites and test versions is minimal, as the coefficients of the regression models for the C3t time scores and task scores were significantly higher than one-hot encoded site and test version variables. It is important to note however that this is only preliminary evidence. Thus, the C3ts site-to-site and test version reliability should be investigated further by a dedicated study.

2.6 Limitations

There are four main limitations to this study. First, the analysed C3t scores are not all the scores the C3t collects. As such, whilst an argument can be made to significantly reduce the complexity of the BTT and CTT, similar arguments cannot be made for any of the other tasks.

The rationale for only looking at the BTT and CTT was that the amount of data available was significantly larger for these tasks than the other transfer tasks (DTT and TTT). For this same reason, the BTT and CTT are the only tasks looked at in chapter 4 which deals with the C3ts instrumentation. Additionally, as this thesis is concerned with assessing relationship between the C3t and movement disorders the three baseline tasks (BVT, CVT, and BAT) were not included as these tasks contain no movement component. Similar research should however be conducted for these other C3t tasks.

The second limitation of this study was the lack of a sufficient sample size for estimating the effect sizes of the C3t and clinical measurements over time. The available dataset was too small for any reliable conclusions to be drawn. Additionally, the abnormally low progression in the clinical measures means that whilst there is some evidence that the C3t scores do not change when the clinical measures do not, there is no evidence the C3t scores are anchored to change in those same clinical measurements. As such, to truly understand whether the C3t is anchored to progression the CUHDRS, UHDRS-TMS, and PIN_{HD} this analysis will need to be re-conducted using a larger sample size in a cohort that does show progression over time.

The third limitation was the lack of exploration into practice effects both between C3t tasks and between multiple C3t tests. Any potential practice effects are important to understand if the C3t is to be used in longitudinal studies, especially if data is to be collected regularly. It is also important to understand if practice effects exist between the tasks (e.g., if taking the BTT before the CTT improves participant performance on the CTT) as this would have further implications for which C3t tasks should be retained.

The fourth and final limitation was the insufficient sample size for building regression models and effect sizes for PIN_{HD}. Regarding regression models, as the C3t is highly correlated with PIN_{HD} it is likely suitable for estimating it with a reasonable degree of accuracy. If regression models can be built using the C3t to estimate PIN_{HD} this could allow for regular approximation of the measure which would be valuable to clinical research. However, with only 16 pre-manifest and prodromal participants available to this study it was not felt appropriate to conduct regression analysis. Regarding effect sizes, as has been repeatedly stated throughout this chapter understanding whether the C3t scores are anchored to changes in accepted clinical measurements is key to showing evidence of the C3ts validity. Understanding whether the C3t is mirrors changes in PIN_{HD} is particularly important that PIN_{HD} is designed for use in in pre-manifest and prodromal populations, typically considered a hard stage of HD to track progression in.

2.7 Conclusions and future work

The data presented here suggests that the BTT and CTT tasks of the C3t are strongly linked to measures of general disease progression and has confirmed the relationship previously reported with the UHDRS-TMS. However, the lack of an observable relationship between the C3t scores and chorea suggests that, in its non-instrumented form, the C3t is not sensitive to chorea in HD. Additionally, this study provides initial evidence that the site the C3t was conducted in and the test version that was used has a minimal impact on C3t performance, as the site/test version a C3t instance came from had virtually no impact on regression model performance. The C3t's reliability between sites and test versions does however need further investigation by a dedicated study as this question was not studied in-depth here. Similarly future work should seek to understand whether any practice effect is present both between C3t tasks and between C3t tests.

It appears the C3t can be substantially simplified by removing all scores other than the time taken scores from the BTT and CTT tasks. Whether the time taken scores from both tasks need to be retained or not will require further investigation in a larger cohort with a focus on replicating the findings presented here and determining whether for the purpose of symptom estimation and clinical evaluation both tasks are required. Additionally, future research should investigate the relationship between the time taken scores. It is important to understand why some participants perform more slowly on one task than the other at extreme ends of the symptom severity spectrum. Additionally, the underlying task protocols for one or both of the BTT and CTT may need to be updated, as the difference in performance between the tasks was not linked to any clinical measure, suggesting they are equally difficult.

Given the simplicity of the C3t following the removal of the additional scores, and the high degree of predictive accuracy the scores produced for the CUHDRS, future the C3t into homes and regularly collect data in a longitudinal study. The resultant dataset should be used to explore the stability of the C3t over time. Additionally, the dataset could be used to estimate the CUHDRS and study how an estimation of that measure behaves over time when measures at highly regular intervals. Future work may also wish to explore whether the C3t can be used to train regression models suitable for estimating PIN_{HD}.

In the context of this thesis the primary finding of this study is that in its current form the C3t is not sensitive to chorea. As such, the non-instrumented C3t cannot be used as a baseline to judge the performance of the instrumented C3t for predicting chorea severity in HD. This necessitates the identification of an alternative baseline against which to judge the relationship of the instrumented C3t with chorea. The natural choice of an alternative baseline is the work conducted by Reilmann *et al.*, (2011), as is discussed in chapter 4.

Chapter 3: Developing a remote data collection platform for the C3t

3.1 Chapter Overview

Due to changing requirements for data collection and data analysis in medical research electronic case report forms are rapidly replacing paper-based case report forms, and are in many cases preferable (Franklin, Guidry and Brinkley, 2011). One of the reasons for this shift is the type of data being collected during medical research is changing, with sensor data being increasingly used to provide enhanced insights into diseases. Whilst traditional clinical assessment data (e.g., demographic information, simple assessment scores like time taken or number correct answers, etc) sensor data is ill-suited for recording using paper-based methods. Another factor driving the shift from paper to electronic collection methods is as most data is now shared and analysed electronically. As such, collecting data using paper-based methods can result in wasted effort as it will usually need to be digitised at some point anyway. This is especially true when data is to be shared between multiple disparate sites. Due to the increasing importance in electronic data collection methods medical research projects can benefit from, and can often outright require, software systems designed to support their collection, transmission, and storage.

During this project, an opportunity arose to embed the C3t and accelerometer collection protocol into multiple clinical studies. This necessitated the production of an electronic data collection solution suitable for collecting C3t data, ensuring its interoperability with accelerometer data, and its transmission & electronic storage.

It is important to note that an app had already been developed prior to this project commencing for the MBT, the original version of the C3t. The MBT was however substantially revised, and the use cases updated (e.g., the need for remote storage, interoperability with accelerometer recordings) leading to the old app to be significantly refined and additional software systems constructed over the course of this project.

The full data collection system for the C3t that was constructed and embedded into clinical studies is referred to throughout as the Remote Data Collection Platform (RDCP). As the construction of the RDCP formed a significant part of this project, is the sole reason the analysis presented in chapter 4 was possible, and the increasingly important place such systems play in research, the RDCP is presented in this chapter.

The overall goal of this chapter is to present a full case study for the design, development, and usability analysis of an RDCP, ultimately providing a series of recommendations and warnings to help guide future research which needs to make use of similar systems.

As such, this chapter has four objectives are addressed in turn throughout the following sections.

Objective 1: Provide an overview of paper-based and electronic-based data collection methods and outline why electronic data collection methods are becoming more prevalent.

Objective 2: Outline the use cases and operational requirements of the RDCP

Objective 3: Detail how the use cases and operational requirements of the RDCP were met

Objective 4: Critically evaluate the performance of the RDCP, in order to provide recommendations for future studies

Please note, the total code base for the C3t app alone is in excess of 10,000 lines of code equating to approximately 208 A4 pages. As such for brevity, implemented code is not included in this chapter nor in the appendix. It can and will be made available upon request via a remote GitHub repository.

3.2 Introduction

In recent years it has become recognised that data is the primary asset of medical research, both in commercial biopharmaceutical enterprises and in scholarly research (Lu and Su, 2010). The importance of data is not limited to medical research however, data drives all areas of research forward; it is the basis on which hypotheses are constructed, the method used to test those hypotheses and the indicator of promising avenues to explore. Data holds just as much importance in scientific fields where hypotheses testing is not the de-facto standard. Regardless of the study area in question, data is the bedrock on which scientific advancement is built.

Given the place of data in scientific enquiry the methods by which it is collected are also important. In medical research data has historically been collected using Case Report Forms (CRFs) – paper-based forms with the fields necessary to record pertinent clinical data that will be used in subsequent analyses (Latha, Bellary and Krishnankutty, 2014). Their structured nature allows for data to be collected in a standardised manner and can include detailed instructions for the clinicians completing them. CRFs can be easily replicated, are cost effective and simple to construct and update.

Medical research methodology has however been undergoing significant change in the last decade, with paper-based CRFs are becoming insufficient for the requirements of modern studies (Lu and Su, 2010).

One of the main limitations of paper CRFs is their ability to collect large amounts of high-quality data from multiple study sites. The process of manually writing potentially large quantities of study & assessment results and can easily result in errors (Nahm, Pieper and Cunningham, 2008). The likelihood of errors occurring will naturally increase as study size and the complexity of the data being recorded also increases. As data is almost always analysed digitally, and physical CRFs are rarely shared between different study sites, the data recorded will need to be entered again electronically either by the clinician or supporting staff members, needlessly increasing workload (Le Jeannic *et al.*, 2014).

This problem is exacerbated when researchers wish to apply machine learning techniques to analyse collected data. Such techniques, whilst prevalent in medical literature, require significantly higher sample sizes than traditional medical research methods (Deo, 2015). It is common practice in medical research to bolster the sample size of studies by having study data be collected at numerous sites. Therefore, as paper CRFs are ill suited to the robust collection of large data sets, particularly across multiple study sites, alternative data collection methods are desirable in such studies.

Whilst paper based CRFs are capable, if not necessarily optimal, for collecting traditional medical research data in large-scale studies they are fundamentally unsuitable for collecting data recorded using modern digital technology. As covered in section 1.4.3, instrumented assessments, which fuse sensor technology with clinical assessments to provide enhanced insights, have been widely applied in medical research and are steadily gaining popularity (Hasan *et al.*, 2017; Zampogna *et al.*, 2020). The outputs from commonly used sensors (such as accelerometers, IMUs, EMGs, etc) are fundamentally incompatible with traditional paper-based data collection. Figure 10 shows a sample output from an accelerometer, making it clear to see why such data is not suited for recording using paper CRFs.

```
2017-05-31 12:11:19:030,-0.7430,-0.4505,-1.0919,0,0,35.9
2017-05-31 12:11:19:040,-0.4118,-0.1556,-0.9407,0,0,35.9
2017-05-31 12:11:19:050,-0.1634,-0.1835,-0.9089,0,0,35.9
2017-05-31 12:11:19:060,-0.0925,-0.1317,-0.8134,0,0,35.9
2017-05-31 12:11:19:070,-0.1792,0.0954,-0.8452,0,0,35.9
2017-05-31 12:11:19:080,-0.1792,0.1632,-0.9129,0,0,35.9
2017-05-31 12:11:19:090,-0.1634,0.2588,-0.9805,0,0,35.9
2017-05-31 12:11:19:100,-0.2147,0.4182,-1.0243,0,0,35.9
2017-05-31 12:11:19:110,-0.1792,0.5019,-1.0601,0,0,35.9
2017-05-31 12:11:19:120,-0.2699,0.3943,-1.1595,0,0,35.9
2017-05-31 12:11:19:130,-0.3645,0.2549,-1.2351,0,0,35.9
2017-05-31 12:11:19:140,-0.3803,0.1632,-1.2471,0,0,35.9
2017-05-31 12:11:19:150,-0.3724,0.0397,-1.1874,0,0,35.9
2017-05-31 12:11:19:160,-0.2817,0.0476,-1.1078,0,0,35.9
2017-05-31 12:11:19:170,-0.2383,0.0954,-1.0601,0,0,35.9
2017-05-31 12:11:19:180,-0.2896,0.0795,-1.0123,0,0,35.9
```

Figure 10: Example comma separated value output file of a GeneActiv accelerometer (ActivInsight; UK). The recording was a little under 14 minutes long with a recording frequency of 100Hz (1 sample every 10 milliseconds, 100 readings per second), resulting in approximately 84000 lines of recorded data.

Due to the shortfalls of paper CRFs in the evolving medical research landscape, there has been a consistent push in recent years towards electronic CRFs (eCRFs). The main concept behind eCRFs is to replace paper-based data collection with electronic collection using a PC, tablet or mobile. eCRFs address many of the shortcomings of paper CRFs by allowing for automatic data checking and automated calculations of derived variables leading to increased data quality and easy integration into multi-centre study databases (Thwin *et al.*, 2007; Le Jeannic *et al.*, 2014).

Due to eCRFs storing and transmitting clinical data electronically, it is important robust standards are met when developing and operating such systems (National Institute for Health Research, 2014). These standards can generally be split up into two types - infrastructure requirements and functional requirements. Infrastructure requirements are based around the physical hardware used to host the system (e.g., data retention systems, web-based collection software, etc) as well as how those systems should be managed. They include stipulations such as all servers being kept in secure rooms, direct electronic access being restricted to system administrators, and the development of a data recovery plan in case of infrastructure failure. Functional requirements are based around how the systems should be used in practice (e.g., software setup, training requirements, data validation). Examples of functional requirements include the procedure for randomising data, audit trails for noting why/when/how/by whom data were entered, data encryption standards, and procedures for the importing/exporting of data to/from a study. These standards exist to both safeguard confidential patient data, as well as to help ensure robust data collection procedures are followed, ultimately heightening data and so study quality.

Despite the additional overhead eCRFs can generate, both in terms of the need to develop them as well as the numerous standard they should adhere to, a comparison review of eCRFs and paper CRFs found that eCRFs were actually typically more efficient than their paper-based counterparts. Le

Jeannic *et al.*, (2014) found that they were cheaper to develop per patient, better suited to multi-centre trials with large sample sizes and generally better accepted by key study stakeholders including primary investigators, research associates and data managers.

In their simplest form, an eCRF need only be an application suitable for collecting & storing simple sets of data (e.g., questionnaire-style data) on a single device. The true benefit of eCRFs however comes into play when more complex datasets are being collected, as is the case with the C3t. Such assessments have their own internal rules regarding test procedure that can be readily enforced using more advanced software, but which simpler solutions do not necessarily allow for. For example, the C3t requires tasks be taken in a specific order, which is simple to enforce using a custom software application. Additionally, if sensor data is to be collected during an assessment it may be the case that additional functionality is required, such as the synchronisation of task and sensor timestamps, necessitating bespoke software solutions that form a part of the eCRF.

As operational complexity increases, additional support (or 'backend') systems also become desirable for supporting data collection. A standard requirement of such support software is the existence of secure database clinicians can input data to, and the facility for researchers and analysts to subsequently draw data from during analysis. A complete software solution might include a software application to facilitate in-clinic data collection, any associated systems necessary to incorporate sensory data, and a backend system to support the transmission and secure storage of recorded data from separate study sites.

Similar to eCRFs, complete software solutions are becoming more and more common in clinical research. Such software solutions can however be difficult for researchers to realise (Franklin, Guidry and Brinkley, 2011), likely due in part to the relevant skillset (i.e., software design & construction) not being a part of every researchers skillset. Nevertheless, numerous such software solutions to aid research have been developed previously (Avidan, Weissman and Sprung, 2005; Romano *et al.*, 2007; Gao *et al.*, 2008; Harris *et al.*, 2009; Franklin, Guidry and Brinkley, 2011; Hiden *et al.*, 2013; Weeks *et al.*, 2013; Herrick *et al.*, 2016). A recent review noted the various benefits such solutions can provide (from the perspective of clinical trials) ranging from enhanced data quality, to ensuring regulatory procedures are followed (e.g., audit trails), to facilitating data transfer and storage (Gazali, Kaur and Singh, 2017) .

In general, the incorporation of eCRFs, mobile technology, and associated software systems is of such relevance to the current research landscape independent advisory groups, such as Clinical Trials Transformation Initiative (CTTI), have been created (Bakker *et al.*, 2019). Given the complexities of working with mobile technology and their requisite software systems, such groups aim to provide information to researchers regarding best practices for the collection of such data (Initiative, 2018). Of particular relevance to this thesis is such groups focus on mobile technologies (e.g., wearable sensors) which are used in instrumented assessments.

A review of mobile technologies used for clinical research by the CTTI found 275 articles which in some way incorporate wearable, ingestible, implantable or otherwise mobile technology into the study protocol (Bakker *et al.*, 2019). Of the articles reviewed by Bakker *et al.,* 67% of the utilised technology were wearables such as the accelerometers used in chapter 4. Bakker *et al.* also note the advantages mobile technology (and its associated software systems) can provide to medical research, including less frequent study visits, increased measurement precision (one of the primary arguments for instrumented assessments), and capturing data regularly allowing for enhanced analysis. Bakker *et al.,* state that a recommendation by the CTTI is that investigators small-scale studies using mobile technology before launching larger trials, in order to understand sensor usability, develop algorithms, and expose any flaws in the developed system. This is particularly notable here as this is essentially what this entire thesis is for the instrumented C3t, with this chapter serving as the report of the pilot study for the data collection system implemented to facilitate the instrumented C3t. Overall, Bakker *et al.*, highlight the increasing importance of such systems and the need for clinicians, engineers, and other study stakeholders to work together to provide adequate solutions and so continue to drive forward research.

This study was fortunate enough, as mentioned in chapter 1, to have the opportunity to develop an eCRF and associated subsystems to facilitate the embedding of the C3t and accelerometers into a series of clinical studies. The rationale for this was similar to that covered so far – the C3t is well suited to being collected electronically, the accelerometers need to be tied together with the C3t data, and the data needs to be collected from multiple study sites and safely stored ready for analysis.

The developed subsystems are collectively referred to here as the C3ts' Remote Data Collection Platform (RDCP). Due to the significant time investment to construct the RDCP and the growing importance of such systems in medical research it was felt that the inclusion of a chapter discussing the system in this thesis was pertinent. Hence, this chapter details the process of developing an RDCP for the C3t, with the aim of showcasing the specific development considerations, final design, production pipeline, and the lessons which were learned from its deployment.

This chapter is structured in a different manner to the other study-based chapters of this thesis. First, a description of the required system is given. Second, the design decisions and technical specification of the developed system are detailed. Finally, the results of its deployment and the lessons learned from it are discussed. It should be noted that recommended eCRF system standards were adhered to during the development of the RDCP. They were however enforced by the data team at Cardiff University's Centre for Trials Research. As such, as well as for brevity, the specifics of how these standards were followed are not discussed in this chapter.

3.3 RDCP use cases, data model, and flow

3.3.1 Overview

Part 1 of this chapter gives a high-level overview of the RDCP spread across three sections. First, the system use cases are described which detail in a high-level fashion what the RDCP will need to be capable of doing. Second, the high-level data models are detailed, which describe what data the RDCP will need to be capable of recording, transmitting, and storing. Third, the system flow for each of the use cases is defined, which illustrates how the different components of the RDCP will communicate and the interactions that will take place when fulfilling each of said use cases.

It should be noted that the system use cases were defined working in conjunction with the original designer of the C3t, Dr Susanne Clinch (Clinch, 2017b). Outside of this collaboration however, no focus groups were run with other stake holders or end users. This is recognised as a limitation of the developed RDCP (in particular the C3t app portion) and as such is discussed in section 4.6.

3.3.2 System use cases

A broad overview of the RDCP requirements were given in section 1.5. The outline notes that there are two core software components, the C3t app and the database backend. As will be discussed in section 3.4, an additional software artefact will be developed as well to facilitate interoperation of the C3t app and accelerometers. The core components however are the C3t app and associated database backend. The first step in designing the RDCP system will be determining the functions these two components need to be able to perform.

In software engineering, the operational requirements of systems are commonly represented in the first instance using use case diagrams (Gemino and Parker, 2009). Use-case diagrams represent the interactions with a system required by different users (also known as 'actors'), where a user can either be a person or another software system. The general concept is to fully describe the high-level functionality of a system based on the interactions users have with it. By doing so it is then possible to design the software components knowing that the core functionality of the system has already been decided upon and finalised, effectively giving software engineers a list of requirements that must be met for the system to be considered complete.

Regarding the RDCP presented here, there are three users/actors – the clinician, the researcher, and the data manager. The clinician needs to be able to use the C3t app to collect C3t data, link that data to a given participant, synchronise the task timestamps with the accelerometers, and store everything in a remote database. The researcher needs to be able to access the data stored in the database, retrieve the data they want to analyse and link a given participants data to other data sources (e.g., UHDRS results) via id codes and demographic information. The data manager needs to be able to access the data stored in the database and perform CRUD operations as required.

As discussed in section 1.5, CRUD operations are the standard operations required by most software systems, allowing for data to be added, retrieved, modified and removed from database systems (Martin, 1983).

These three users and their respective use cases are shown in Figure 11.

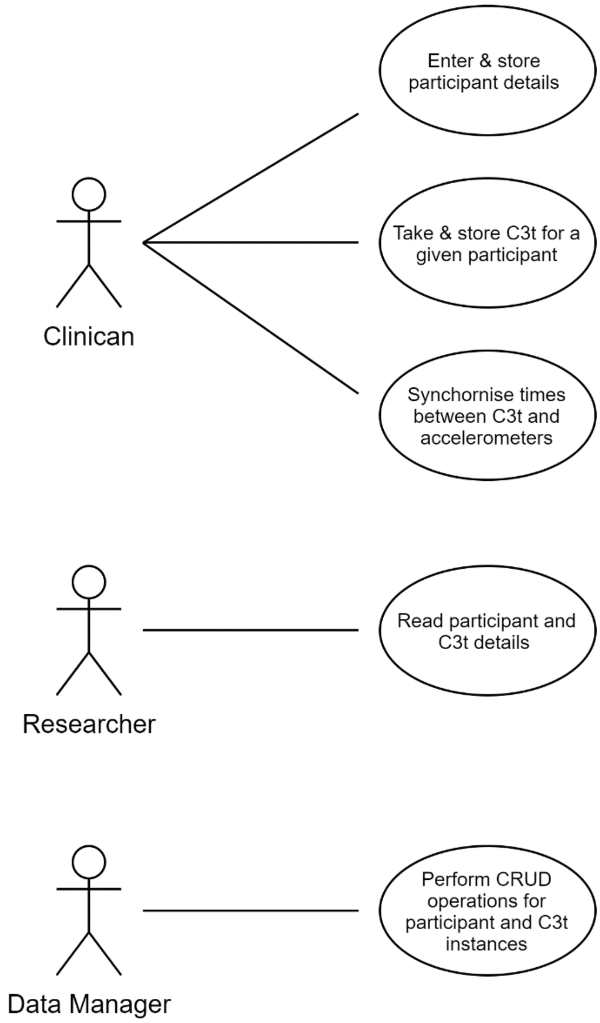


Figure 11: Use case diagram for the three users of the RDCP.

The use cases shown above illustrate that, other than synchronising times between the C3t app and accelerometers, the operations for each user are effectively CRUD operations.

In general, the use cases of the system can be boiled down to the following.

1) CRUD operations for participants
2) CRUD operations for C3t instances
3) Syncing a C3t instance with the PC used to configure the accelerometers

The first and second use cases will allow for participant demographic information to be added and for individual participants to have one or more C3t instances recorded and linked to them in a remote database. The third use case allows for accelerometer data to be collected alongside the C3t and linked to the individual tasks. Whilst the sensors themselves are not collected by the RDCP, the ability to relate specific C3t tasks with the correct accelerometer readings is crucial analysing the two datasets together. If the timestamps are not synchronised accelerometer readings cannot be reliably related together as the clock times of the sensors and C3t app may differ. As such, the RDCP needs to provide a mechanism by which the timestamps of the C3t tasks and accelerometer readings can be synchronised.

3.3.3 Data model

The use cases presented in the preceding subsection illustrate that two distinct chunks of data will need to be recorded, transmitted, and stored in the system – participant details and C3t test results.

As was introduced in section 1.5, a widely used concept used when real world phenomena need to be encapsulated in software is Object Oriented Programming (OOP). The general concept of OOP is to encode the properties/attributes of a real-world 'thing' in a set structure (Oriented, Programming and Oo, 2001). The object may then also have functions/methods which in some way make use of or alter those properties, possibly communicating with other objects of the same or different type. Two common examples of objects are a Bank Account object, which might have a property *balance* and a method *make payment* and a Text Message object which might have a property *content* and a method *send message*. As a side note, objects defined in OOP are usually capitalised, as is the case with the Bank Account and Text Message object, as they are proper nouns.

As the RDCP is primarily based around the recording, transference, and storage of participant and C3t data, two primary objects can be defined – a Participant object which encapsulates the characteristics of an individual participant, and a C3t object which encapsulates an instance of the C3t. These objects will need to be in created by the C3t app as objects and then transmitted and stored in the database backend. The specific design of each object/table for the system components is discussed in section 3.4. The remainder of this section is a high-level breakdown of specifically what data will need to be included in the object/tables across both components however they are ultimately made up.

First however, it is necessary to discuss how time is stored in software systems. The reason for this is that time plays a particularly important role in the RDCP's data model, being one of the main components of the C3t and integral to properly linking C3t data with accelerometer data. As such, how time is represented in software will be briefly introduced before defining the data model requirements.

3.3.3.1 Data model: representing time in software

Measuring time is crucial for both the C3t, instrumented C3t and RDCP. In the C3t, the time taken to complete each task is taken is one of the primary measures taken during the test. In the instrumented C3t, individual tasks need to be synchronised with accelerometer signals using timestamps. In the RDCP the date C3t instances were taken needs to be recorded as do participant birth dates.

Unfortunately, time can be complicated to handle in software. One of the complexities of handling time in computer systems and databases is due to time zones. To illustrate this, suppose it is 09:00 on 04/07/2020 and from the list below it was necessary to determine which C3t was most recently recorded.

- 03/07/2020, 09:00; New York (US)
- 03/07/2020, 11:00; London (UK)
- 03/07/2020, 13:00; Tokyo (Japan)

Looking at the times and dates alone it would appear that the C3t taken at 13:00 on 03/07/2020 would be the most recent test. However, New York is in UTC-9, London in UTC+0 and Tokyo in UTC+9. As such, test taken at 09:00 in New York would be the most recent test, as 09:00 in New York corresponds to 14:00 in the UK and 23:00 in Japan.

This problem becomes more complicated when daylight saving adjustments are made. Suppose again there was a list of C3t times as follows and again you wanted to pick the most recent test.

- 03/07/2020, 11:30; London (UK)
- 03/07/2020, 11:00; Reykjavik (Iceland)

As the UK and Iceland share the same time zone (UTC+0) the above list suggests that the test taken at 11:30 in London would be the most recent one. However, the UK observes daylight savings and Iceland does not. As such during the summer, 11:00 in Iceland corresponds to 12:00 in the UK, and so the test taken 11:30 in the UK occurred before the test taken at 11:00 in Iceland, despite them being in the same time zone. Whilst this example is slightly contrived it illustrates why handling time in software systems is not as straightforward as it may initially appear to be.

In order to handle this problem in software there are two choices, either the logic for all time differences including cultural, political, and geographical variations must be encoded in the system or the problem needs to be somehow avoided. Typically, the latter approach is taken. ISO 8601 (originally published in 1988) sets out the concept of a standard time epoch, a static point in time from which computer systems calculate their current time relative to (International Standards Organization, 2004).

There are a number of different epochs used, however arguably the most common is the Unix epoch, also called a Unix timestamp, which defines time as the number of seconds or milliseconds relative to 00:00 on January 1st, 1970 UTC+0 (International Standards Organization, 2004). Unix timestamps are represented as real integers where negative values are before the epoch and positive values are after the epoch. Unix timestamps are used throughout all RDCP software components to represent data and time. Three examples of human-readable timestamps and their corresponding Unix timestamps are shown in Table 23.

Table 23: Examples of human-readable timestamps and their corresponding Unix timestamps in millisecond format

Time and date	Unix Timestamp (milliseconds)
00:00, December 31st, 1969	-86400000
00:00, January 1st, 1970	0
00:00, January 2nd, 1970	86400000

3.3.3.2 Data model: Participant data model

The first part of the data model is the participant data model. The participant data model should be sufficient for uniquely identifying a participant within the study without using a name, such that individual C3t instances can be related to a given participant. Participant details including date of birth (represented as a Unix timestamp) and gender will be needed for demographic purposes. The dominant hand will also need to be recorded such that sensor data can be compared in the same manner across people whose dominant hands differ. Overall, four fields of participant data are required as follows.

- Study unique identifier
- Gender
- Date of birth
- Dominant hand

3.3.3.3 Data model: C3t data

The C3t data model will need to capture all measures listed in the C3t manual across all six tasks as well as several additional fields, making it significantly larger than the participant data model.

Three additional bits of C3t information for a specific C3t instance which is not described in the manual will be also needed as follows.

- Study unique identifier of participant a C3t instance belongs to
- The date a given C3t instance was taken
- Whether sensor calibration was performed, assuming they are in use

There will be two departures from the items listed in the C3t manual. First, instead of the time each task takes being recorded the timestamp a task started and the timestamp a task finished will be needed instead. Having the start and finish timestamps of a task will allow the sensor data, when available, to be synchronised with a given C3t task. Second, as the various study sites are not all based in countries where English is the primary language, the DTT and BAT tasks (which incorporate a spoken alphabet element) will need to allow for different languages to be used. As such, which language was used will need to be recorded as some of the C3t measures rely on the number of letters in the used alphabet for their calculations.

Table 24 shows each of the recorded and derived measures that will need to be included in the RDCP data models, which tasks they apply to, and the additional fields as appropriate.

Table 24: Recorded and derived measures of the C3t updated to include measures specific for the RDCP system. Values not found in the C3t manual are italicised.

Recorded Measures		
Measure	Description/Equation	Applicable Tasks
Time taken in seconds	The time a task took to complete in seconds (accurate to 2 decimal places)	All tasks (6)
Time started (milli seconds since epoch)	The time a task was started in milliseconds relative to the Unix epoch	All tasks (6)
Time finished (milli seconds since epoch)	The time a task was finished in milliseconds relative to the Unix epoch	All tasks (6)
Rule errors	The number of tokens picked up in the wrong order (maximum 8)	All transfer tasks (3)

Transfer errors	The number of times participants did not correctly transfer tokens between their hands (maximum 8)	All transfer tasks (3)
Dropped tokens	The number of times participants dropped tokens (maximum 8)	All transfer tasks (3)
Correct values	The number of values said in the correct order (maximum 8)	BVT, CVT (2)
Correct letters	The number of letters said in the correct order	BAT, DTT (2)
DTT alphabet time	The time taken to complete the alphabet for the first time	DTT (1)
Alphabet used	The alphabet system used (e.g., English, Welsh)	BAT, DTT (2)
Total recorded measures: 29		
Derived Measures		
Measure	Description/Equation	Applicable Tasks
Transfer task accuracy	$\frac{16 - (rule\ errors + transfer\ errors)}{16} * 100$	All transfer tasks (3)
Transfer task total score	$\frac{8 - dropped\ tokens}{time\ taken\ in\ seconds} * tranfer\ task\ accuracy$	All transfer tasks (3)
Value accuracy	$\frac{correct\ values}{8}$	BVT, CVT (2)
Alphabet accuracy	$\frac{correct\ letters\ (first\ time\ for\ DTT)}{Total\ number\ of\ letters\ in\ alphabet\ system} * 100$	BAT, DTT (2)
Correct letters per second	$\frac{correct\ letters\ (all)}{time\ taken\ in\ seconds}$	BAT, DTT (2)
BTT-CTT time cost	$\frac{CTT\ time\ taken - BTT\ time\ taken}{CTT\ time\ taken}$	Fusion of BTT & CTT (1)
BTT-CTT total score cost	$\frac{CTT\ total\ score\ - BTT\ total\ score}{CTT\ total\ score}$	Fusion of BTT & CTT (1)
CTT-DTT time cost	$\frac{DTT\ time\ taken - CTT\ time\ taken}{DTT\ time\ taken}$	Fusion of CTT & DTT (1)
CTT-DTT total score cost	$\frac{DTT\ total\ score\ - CTT\ total\ score}{DTT\ total\ score}$	Fusion of CTT & DTT (1)
BAT-DTT alphabet cost	$\frac{DTT\ alphabet\ accuracy - BAT\ alphabet\ accuracy}{DTT\ alphabet\ accuracy}$	Fusion of BAT & DTT (1)
Total derived measures: 17		

3.3.4 System Flow

With the system use cases and overarching data model defined, the final high-level description of the system necessary are interaction flow diagrams for the system. These diagrams describe, in an implementation agnostic manner, the flow of operations through the system and the interaction between different system components that will allow the use cases can be fulfilled. The four necessary flows are as follows.

1. Creating participants
2. Displaying/reading current participants
3. Updating/deleting participants
4. Taking a C3t instance
5. Displaying/reading a C3t instance
6. Updating/deleting a C3t instance

Interaction flow diagrams for each of these are shown in the figures below. Creating participants is shown in Figure 12, displaying/reading participant data in Figure 13, updating/deleting participants in Figure 14, taking the C3t in Figure 15, displaying/reading a C3t instance in Figure 16, and updating/deleting a C3t instance in Figure 17. It should be noted that the flow diagram for synchronising the accelerometer and C3t timestamps is included in the fourth interaction flow diagram.

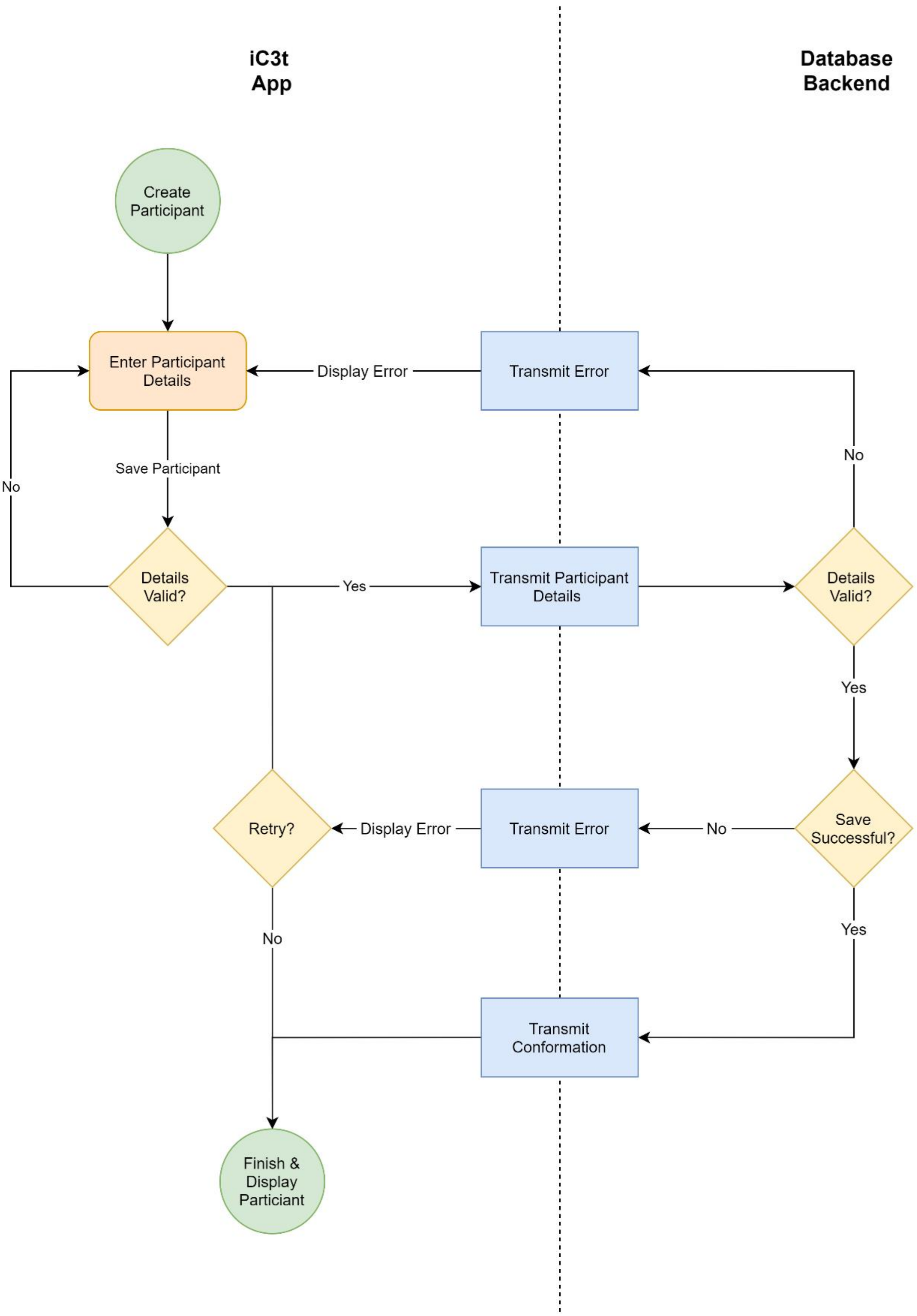


Figure 12: Flow diagram showing the steps for creating a new participant using the C3t app and storing it in the database.

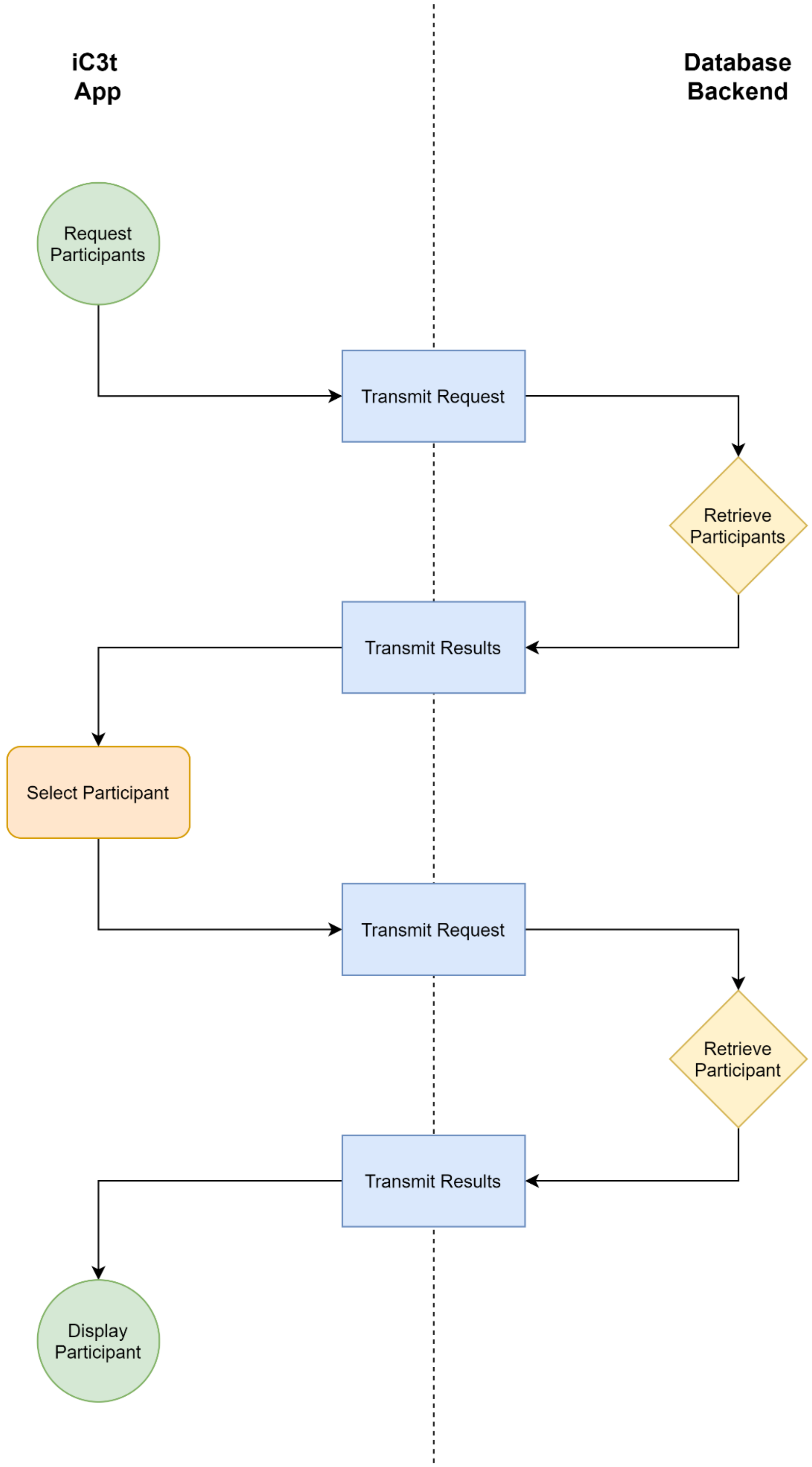


Figure 13: Flow diagram showing the steps for displaying all participants on the C3t app stored in the database before selecting a single participant and showing more details.

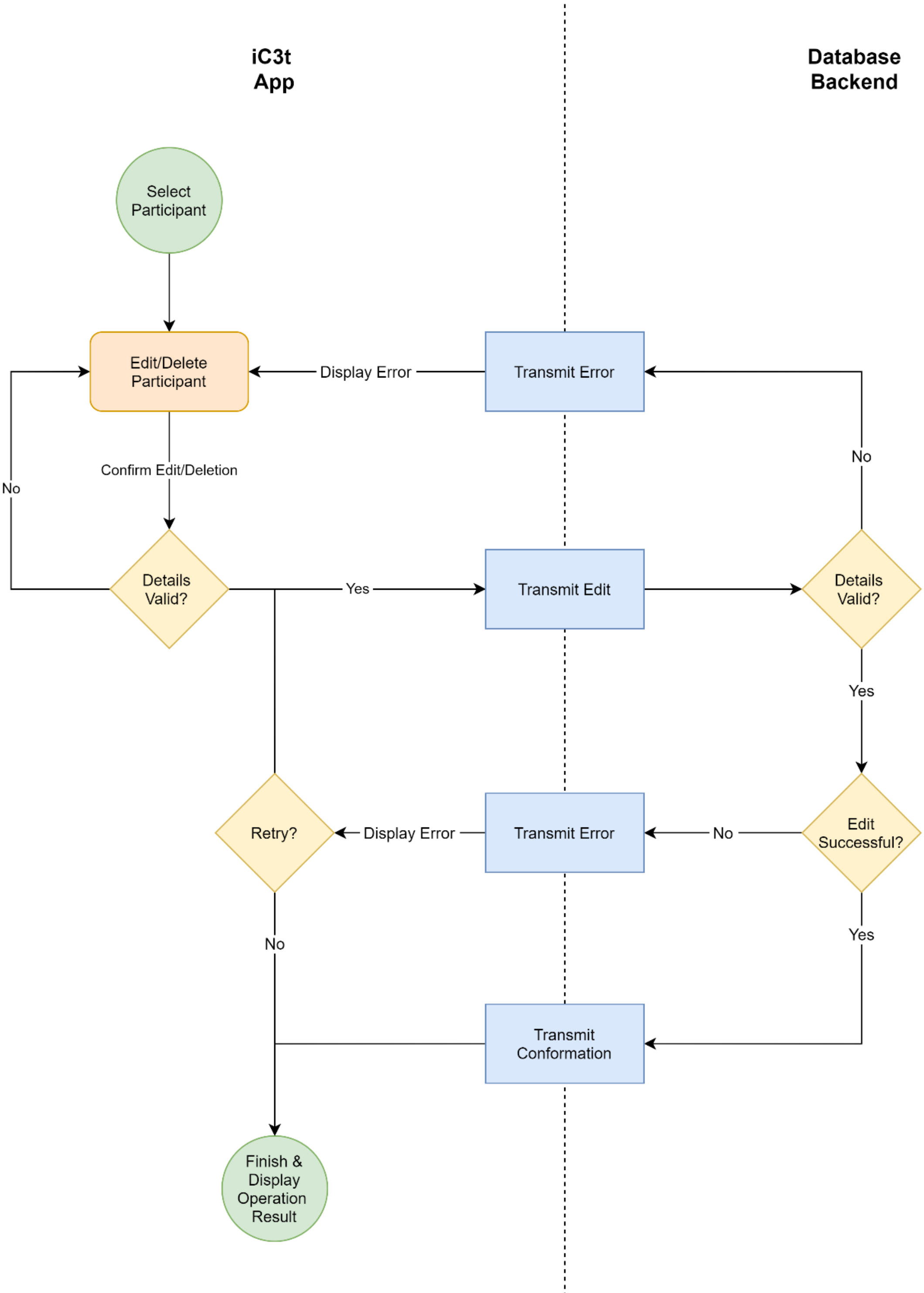


Figure 14: Flow diagram showing the steps for updating/deleting a participant using the C3t app currently stored in the database.

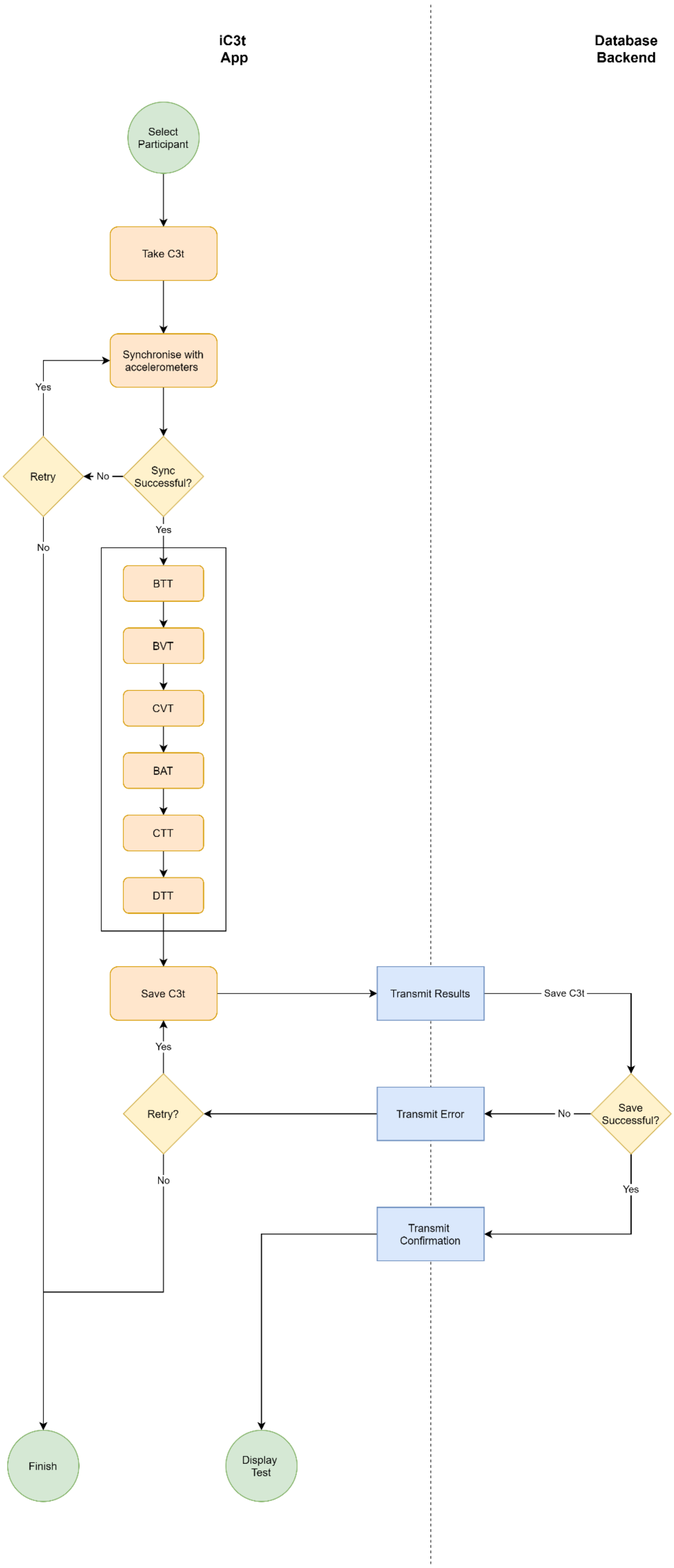


Figure 15: Flow diagram showing the steps for taking the C3t for a given participant using the C3t app and storing it in the database.

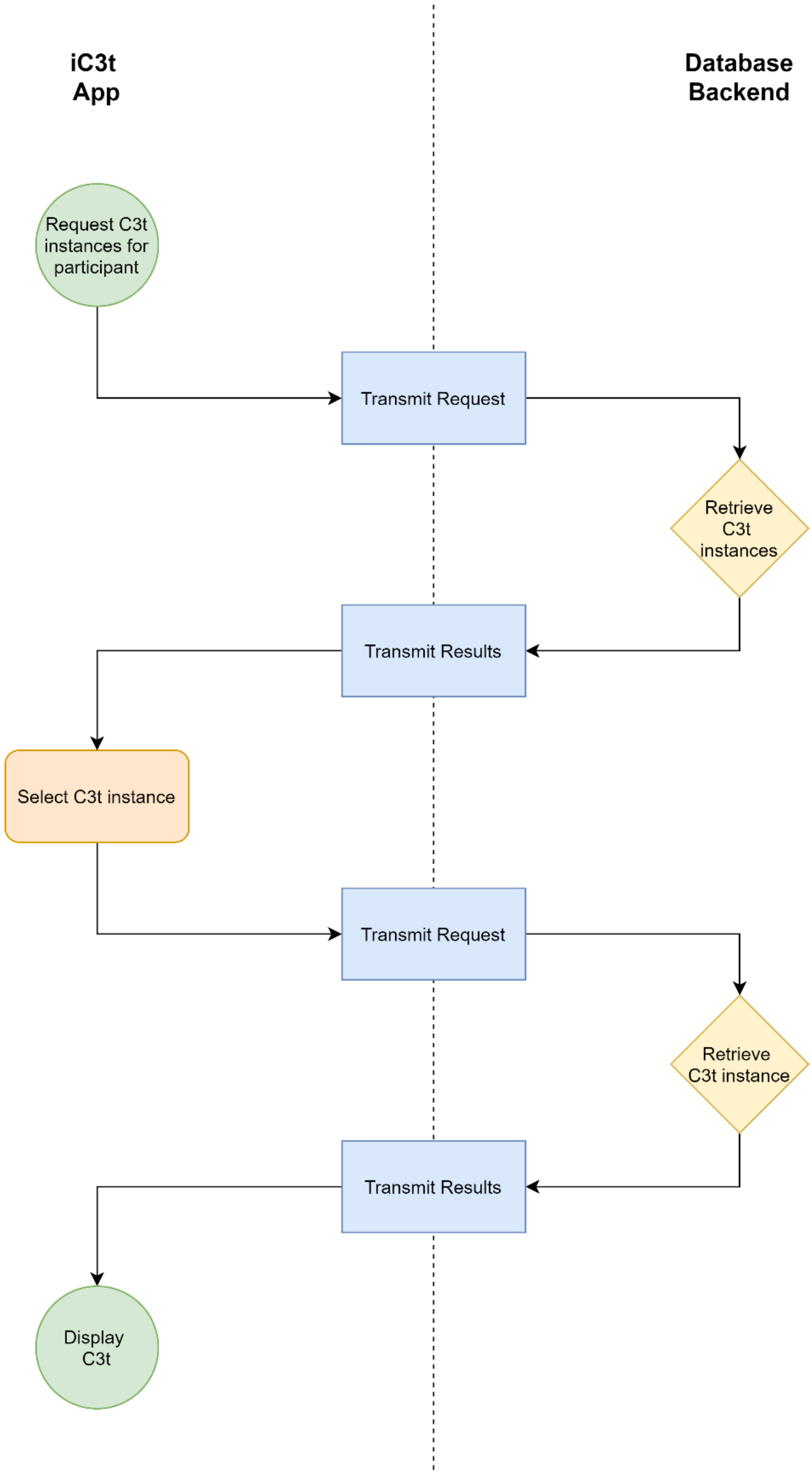


Figure 16: Flow diagram showing the steps for displaying all C3t instances for a particular participant on the C3t app stored in the database before selecting a specific C3t instance and showing the full test results.

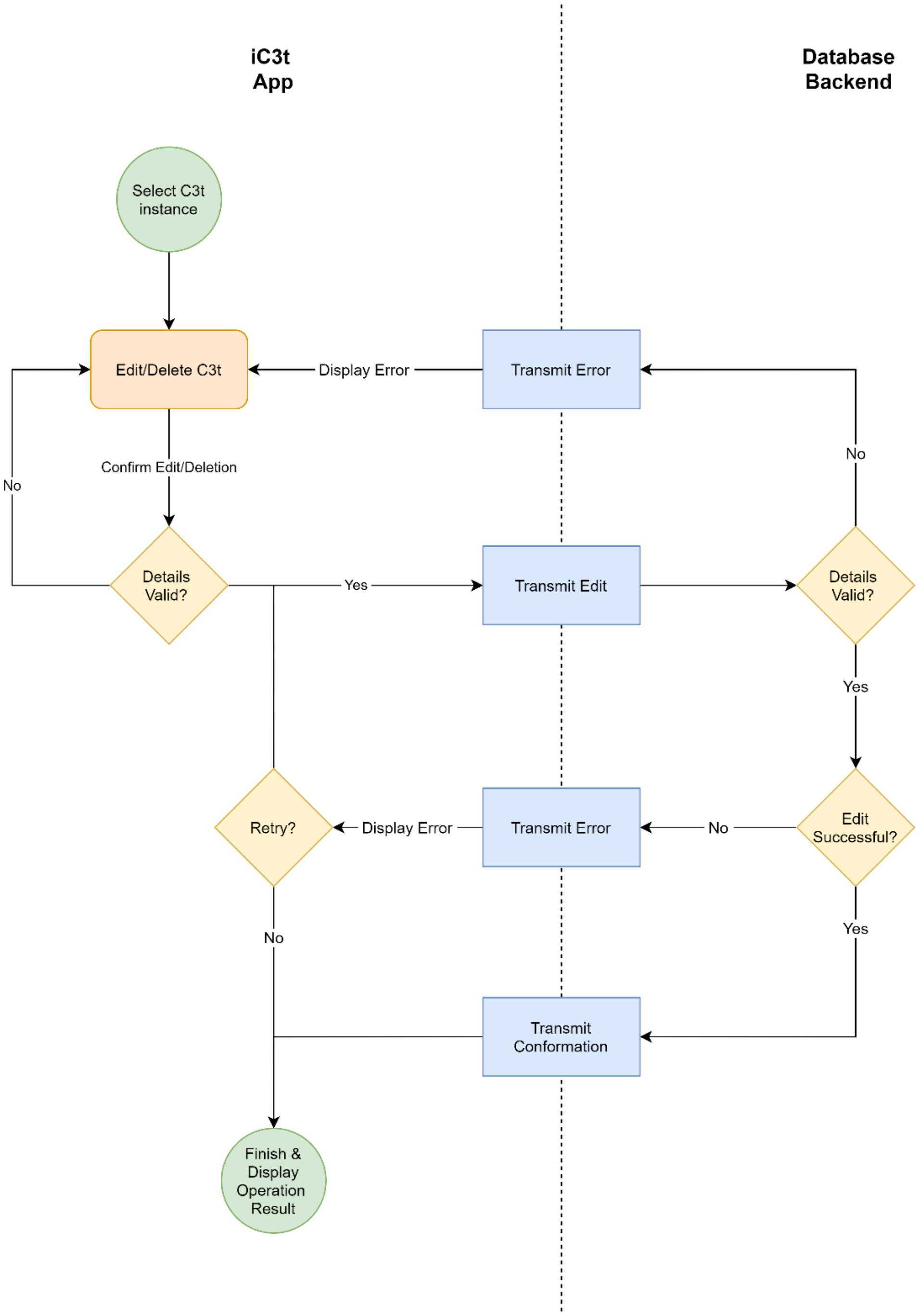


Figure 17: Flow diagram showing the steps for updating/deleting a C3t instance using the C3t app currently stored in the database.

3.4 RDCP technical specifications, design, and implementation

3.4.1 Overview

In the previous section, the high-level requirements of the RDCP were given including the use cases, data model and general system flow.

In this section, the specific implementation details of these requirements will be discussed. The first subsection gives general component-level specifications and functionality descriptions. The second and third subsections then discuss the data models implemented in the C3t app and database backend, respectively. The final subsection shows how each use case was implemented across the RDCP as a whole with accompanying screenshots and process descriptions.

3.4.2 General component specification

Three software components were constructed to fulfil the requirements outlined in section 3.3 as follows.

1) The C3t app
2) A PC-based app
3) The database backend

The C3t app allows clinicians to collect data, view data, and facilitates the app-side of the C3t-accelerometer timestamp synchronisation. The PC app facilitates the PC/accelerometer side of the C3t-accelerometer timestamp synchronisation. The database backend facilitates the storage and retrieval of data collected via the C3t app.

The C3t app was programmed for Android 7.0 (Nougat) using Java 8 but has recently updated to support Android 9 (Pie) using Java 10. The PC app was programmed using Java 8. The database backend was built using Microsoft SQL Server with all communication elements being performed by a series of PHP 7.2 scripts.

The form of the RDCP is shown in Figure 18.

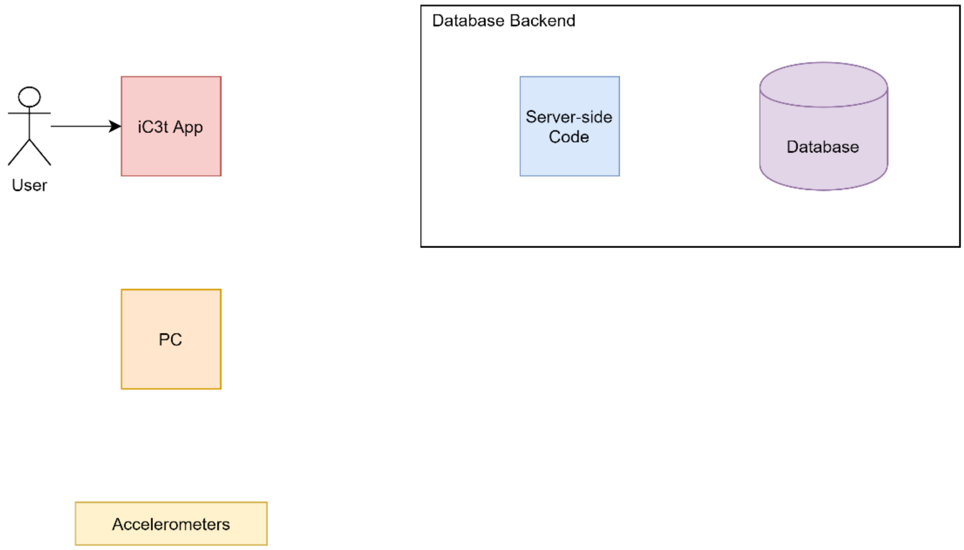


Figure 18: General form of the RDCP showing the flow of communication between connected parts. The three components developed here are the C3t app, PC app, and database backend. Only components which are connected together directly communicate.

3.4.3 C3t app data model

3.4.3.1 C3t app data model overview

The C3t app is structured around two main concepts – participants and C3t instances. These concepts form the basis for the data model used in the C3t app – participants (which are described by their attributes) take C3t instances (which are described by their results). The specific data participants and C3t instances need consist of and thus the C3t app needs to record were listed previously in section 3.3.3.3.

To represent the concepts of participants and C3t instances objects were used (as introduced in section 1.5 and briefly again in section 3.3.3). Ultimately 10 objects were defined – Participant, C3t, Transfer Task, Baseline Task, and one for each of the C3t tasks. High-level descriptions of these objects are shown in Table 25. The class diagram of the C3t app is shown in Figure 19 and details the structure of the C3t app data model including object types & descriptions as well as the logical relationships between them. Generic methods (i.e., getters, setters, and constructors) have been omitted for brevity.

Table 25: High-level descriptions of each object in the C3t app.

<table>
<tr><th>Object Name</th><th>Description</th></tr>
<tr><td>Participant</td><td>Represents the concept of an individual participant, containing attributes suitable for describing the participants details as listed in Data model: Participant data</td></tr>
<tr><td>C3t</td><td>Represents the concept of an entire C3t instance, containing the extra information in the first three bullet points at the start of Data model: C3t data, as well as holding a single instance of each task-specific objects</td></tr>
<tr><td>Transfer Task</td><td>An abstract object that contains the attributes and methods common across the three transfer task objects (BTT, CTT, DTT)</td></tr>
<tr><td>Baseline Task</td><td>An abstract object that contains the attributes and methods common across the three baseline task objects (BVT, CVT, BAT)</td></tr>
<tr><td>BTT</td><td rowspan="6">Task specific objects for each of the 6 C3t tasks, containing the results for each of these tests as well as the functions necessary to compute their derived results</td></tr>
<tr><td>CTT</td></tr>
<tr><td>DTT</td></tr>
<tr><td>BVT</td></tr>
<tr><td>CVT</td></tr>
<tr><td>BAT</td></tr>
</table>

Please note that the Transfer Task and Baseline Task objects do not exist in the current version of the C3t app. In its current form the tasks are each uniquely defined rather than inheriting from two base classes. What is presented in Figure 19 is what the app originally looked like before one of the development challenges (detailed later on in this chapter, see section 3.5) made it necessary to reduce the elegancy of the solution in order to increase the rate at which alterations to the code base could be made. As the development challenge that made this necessary is no longer a factor, the C3t app will be updated to the original form described in Figure 19 in preparation for the end of its support provided by this project. This process is known in software engineering as ‘refactoring’ and is a common step towards the end of a project (McConnell, 2004).

The rationale for including the original & future design rather than what is currently implemented was that the complexity of the solutions description is greatly reduced but the functionality is identical. The goal of this chapter is to give an overview of the RDCP functionality and a discussion regarding its efficacy. As such, it was felt that presenting the more elegant design would reduce complexity for the reader without effecting the accuracy of the description of the systems functionality.

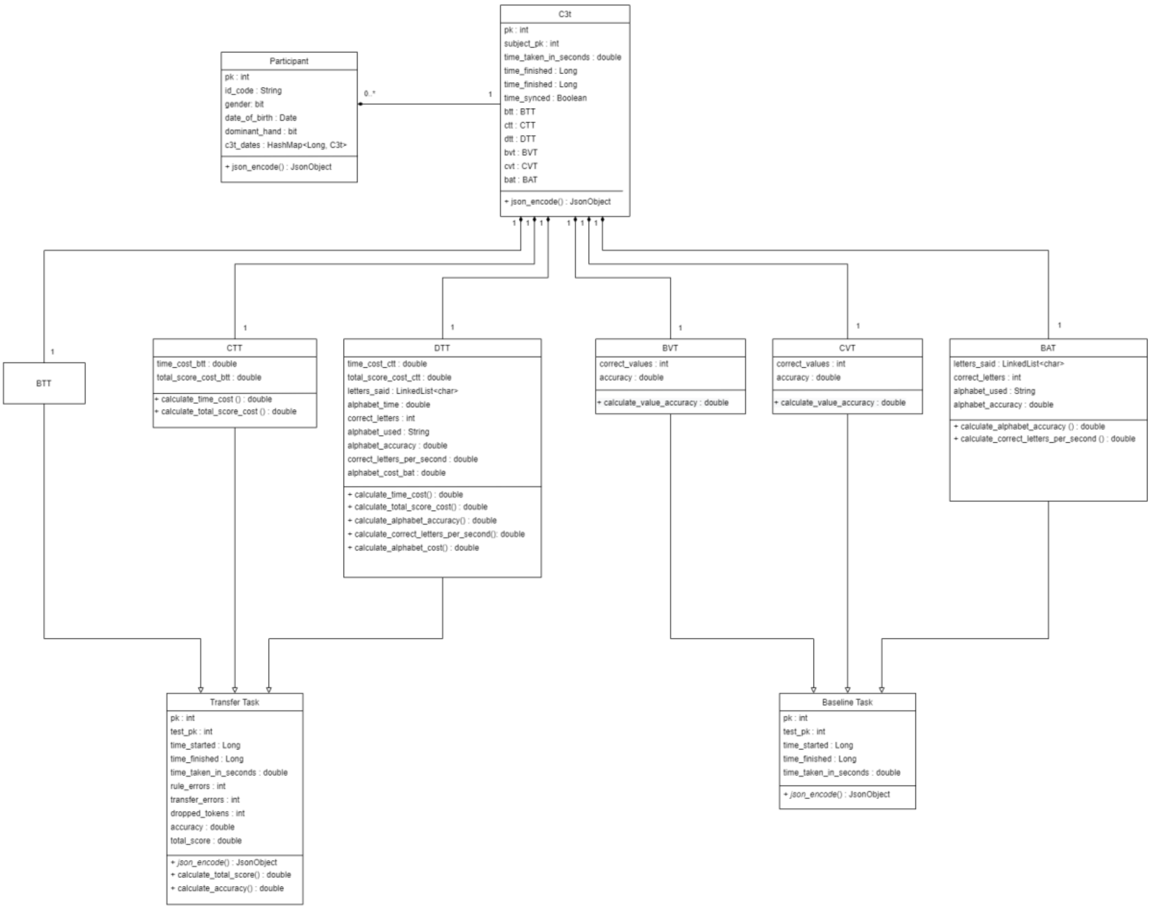


Figure 19: Class diagram showing the C3t app data model. Black diamonds indicate multiplicity (i.e., ownership) with the diamond side indicating a 'has-a' relationship with the plain side. Numerals indicate the degree of multiplicity (e.g., 0.. indicates a 0 to many relationship). White arrows indicate an inheritance relationship, with the arrow-side being the parent class and the blank-side being the child class. The boxes represent a single object. The top text is the objects name, the second row of text are the attributes, and the third row are the object methods. The attributes follow the structure **name : type** and the methods follow the structure **visibility (+) method signature : return type**. Italicized method signatures indicate an abstract method which must be implemented in the inheriting object.*

The following subsections are short natural language descriptions of the details shown in Figure 19 starting with the one method common across all objects, JSON Encode.

3.4.3.2 Json Encode

All objects have a *json_encode* method. This method encodes the object attributes data as a JSON string such that it can be transmitted to the database backend. An example of a BTT instance encoded in JSON is shown in Figure 20.

```
{
    "pk": "1"
    "test_pk": "CON-001",
    "time_started": "1594033200000"
    "time_finished": "1594033260000"
    "time_taken_in_seconds": "60.0"
    "rule_errors": "0"
    "transfer_errors": "0"
    "dropped_tokens" : "0"
    "accuracy": "100.0"
    "total_score" : "100.0"
}
```


Figure 20: Example of BTT results encoded in JSON string.

3.4.3.3 Participant Object

The Participant object contains the attributes shown in Table 26 below. It is related to the C3t object via a 'has-a' relationship, indicated by the black-diamond link between Participant and C3t, which encodes the concept of a participant having taken the C3t some number of times. The number of these relationships can be as few as 0 with no specified upper limit.

The Participant object makes use of the concept of an 'enum'. An enum is a special data type in Java which allows a value to take only a set of pre-defined values. Enums allow for seamless conversion between text and integers, with the position of the enum in its declaration acting as the numeric representation of the enum. For example, the dominant hand enum can be either 'dominant' or non-dominant', with each assigned an integer of 0 or 1, respectively.

Table 26: Description of participant attributes as shown in Figure 19

Name	Type	Description
pk	int (integer)	The primary key of the subject used by the database to uniquely locate a given participant
id_code	String (text)	The study id code for the participant
date_of_birth	Date	The date of birth of the participant encoded as a Java Date object but created and transmitted using the millisecond method described previously

dominant_hand	Enum	The dominant hand of the participant which uses an Enum (left or right)
gender	Enum	The gender of the participant represented as an Enum (male, female, unspecified)
c3t_dates	HashMap<Long, C3t>	A hash-map containing the time each C3t instance the participant has taken was started mapped to the C3t object itself, allowing for the C3t objects to be simply ordered and displayed using the date and time each was taken

3.4.3.4 C3t Object

The C3t object contains the attributes shown in Table 27 below. C3t objects are owned by Participant objects and cannot exist without being associated with a single participant. The C3t object encodes the high-level idea of a single C3t being taken. It has a 'has-a' relationship with each C3t task and must have 1 of each task.

Table 27: Description of C3t attributes as shown in Figure 19

Name	Type	Description
pk	int (integer)	The primary key of the C3t instance used by the database to uniquely locate a given participant
participant_pk	int (integer)	The primary key of the participant used this C3t instance belongs to
time_started	long (real integer)	The time the test was started in milliseconds relative to the Unix epoch

time_finished	long (real integer)	The time the test was finished in milliseconds relative to the Unix epoch
time_taken_in_seconds	double (real decimal place)	The time the whole test took in seconds
time_synced	Boolean (true or false)	Whether or not this test instance was synchronised with a computer such that accelerometer data can be included
BTT CTT DTT BVT CVT BAT	Various Task objects	Individual C3t task objects which are described in the following subsections

3.4.3.5 Transfer Task & BTT Objects

The Transfer Task object contains the attributes and methods shown in Table 28 below. This object is used to aggregate common attributes of the transfer tasks (BTT, CTT & DTT) together into a single object. As is shown in Table 24, the three transfer tasks each contain many of the same scores. Thus, it makes sense to combine them into a single class for the purpose of displaying the data model. The individual transfer tasks 'inherit' the properties of the Transfer Task object, meaning they share its attributes and methods whilst being able to add additional attributes and methods to themselves as needed.

Unlike the other transfer tasks, the BTT object has no additional attributes or methods outside of those described by the Transfer Task object. As such, it is represented in Figure 19 as a blank object and is fully described by Table 28. The BTT object was created as its own method as in the case it requires additional attributes not shared by the other tasks at a later date.

Table 28: Description of Transfer Task attributes and methods as shown in Figure 19

Name	Type	Description
pk	int (integer)	The primary key of the task instance used by the database to uniquely locate the task
test_pk	int (integer)	The primary key of the C3t instance used this task instance belongs to
time_started	long (real integer)	The time the task was started in milliseconds relative to the Unix epoch
time_finished	long (real integer)	The time the task was finished in milliseconds relative to the Unix epoch
time_taken_in_seconds	double (real decimal place)	The time the task took in seconds
rule_errors	int (integer)	The number of rule errors made during the task
transfer_errors	int (integer)	The number of transfer errors made during the task
dropped_tokens	int (integer)	The number of tokens dropped during the task
accuracy	double (real decimal place)	The accuracy of the task
total_score	double (real decimal place)	The total score of the task
calculate_accuracy()	method, returns double (real decimal place)	Calculates the accuracy of the task
calculate_total_score()	method, returns double (real decimal place)	Calculates the total score of the task

3.4.3.6 CTT Object

The CTT object contains the attributes and methods shown in Table 29 below. It inherits from the Transfer Task object and thus contains all the attributes and methods shown in Table 28. Additionally, it adds two additional attributes and two methods used to calculate those attributes. The CTT object encodes the concept of a single CTT instance within a C3t instance.

Table 29: Description of CTT attributes and methods as shown in Figure 19

Name	Type	Description
time_cost_btt	double (real decimal place)	The difference between the time taken to complete the BTT and CTT
total_score_cost_btt	double (real decimal place)	The difference between the total scores of the BTT and CTT tasks
calculate_time_cost()	method, double (real decimal place)	Calculates the difference in time taken to complete the BTT and CTT tasks
calculate_total_score_cost()	method, returns double (real decimal place)	Calculates the difference in total scores between the BTT and CTT tasks

3.4.3.7 DTT Object

The DTT object contains the attributes and methods shown in Table 30 below. It inherits from the Transfer Task object and thus contains all the attributes and methods shown in Table 28. Additionally, it adds nine additional attributes and five methods used to calculate those attributes. The DTT object encodes the concept of a single DTT instance within a C3t instance.

Table 30: Description of the DTT attributes and methods as shown in Figure 19

Name	Type	Description
time_cost_ctt	double (real decimal place)	The difference between the time taken to complete the CTT and DTT
total_score_cost_ctt	double (real decimal place)	The difference between the total scores of the CTT and DTT tasks
letters_said	LinkedList<String> (linked list of Strings)	The order in which letters were said during the DTT

alphabet_time	double (real decimal place)	The time taken in seconds for the participant to complete one cycle through the alphabet
correct_letters	int (integer)	The number of letters said in the correct order
alphabet_used	enum	The alphabet system used
alphabet_accuracy	double (real decimal place)	The number of letters in the correct order over the number of total letters said (first pass only)
correct_letters_per_second	double (real decimal place)	The number of correct letters said per second (entire task)
alphabet_cost_bat	double (real decimal place)	The difference between the alphabet accuracy of the BAT and the alphabet accuracy of the DTT
calculate_time_cost()	method, double (real decimal place)	Calculates the difference in time taken to complete the CTT and DTT tasks
calculate_total_score_cost()	method, returns double (real decimal place)	Calculates the difference in total scores between the CTT and DTT tasks
calculate_alphabet_accuracy()	method, returns double (real decimal place)	Calculates the alphabet accuracy (first pass only)
calculate_correct_letters_per_second()	method, returns double (real decimal place)	Calculates the number of correct letters per second (entire task)
calculate_alphabet_cost()	method, returns double (real decimal place)	Calculates the alphabet cost in terms of accuracy between the BAT and DTT

3.4.3.8 Baseline Task Object

The Baseline Task object contains the attributes and methods shown in Table 31 below. This object performs a similar function for the baseline task that the Transfer Task object does for the transfer tasks by combining common attributes and methods into a single object.

Table 31: Description of the Baseline Task attributes as shown in Figure 19

Name	Type	Description
pk	int (integer)	The primary key of the task instance used by the database to uniquely locate the task
test_pk	int (integer)	The primary key of the C3t instance used this task instance belongs to
time_started	long (real integer)	The time the task was started in milliseconds relative to the Unix epoch
time_finished	long (real integer)	The time the task was finished in milliseconds relative to the Unix epoch
time_taken_in_seconds	double (real decimal place)	The time the task took in seconds

3.4.3.9 BVT & CVT Object

The BVT and CVT objects contain the attributes and methods shown in Table 32 below. The BVT and CVT both inherit from the Baseline Task object and so contain the same attributes and methods described in Table 31. They also have two additional attributes and one additional method as shown below.

Table 32: Description of the BVT and CVT attributes and methods as shown in Figure 19

Name	Type	Description
correct_values	int (integer)	The number of values said in the correct order during the BVT or CVT
accuracy	double (real decimal place)	The number of correct values spoken divided by the total number of values to be spoken
calculate_value_accuracy	method, returns double (real decimal place)	Calculate the accuracy of the task

3.4.3.10 BAT Object

The BAT object contains the attributes and methods shown in Table 33 below. The BAT object inherits from the Baseline Task object and so contains the same attributes and methods described in Table 31. It also has three additional attributes and two additional methods as shown below.

Table 33: Description of the BAT attributes and methods as shown in Figure 19

Name	Type	Description
letters_said	LinkedList<String> (linked list of Strings)	The order in which letters were said during the BAT
correct_letters	int (integer)	The number of letters said in the correct order
alphabet_used	enum	The alphabet system used
alphabet_accuracy	double (real decimal place)	The number of letters in the correct order over the number of total letters said
correct_letters_per_second	double (real decimal place)	The number of correct letters said per second
calculate_alphabet_accuracy()	method, returns double (real decimal place)	Calculates the alphabet accuracy
calculate_correct_letters_per_second()	method, returns double (real decimal place)	Calculates the number of correct letters per second

3.4.4 Database backend data model

The database backend system needs to suitable for recording the participant and C3t data captured by the C3t app. As such, the fields used by the database tables are very similar to the attributes of the C3t app objects, but without the complexities of inheritance or methods. The database does however make use of the concepts of primary and foreign keys. Primary keys are identifiers that are guaranteed by the database management system to be unique for each row in a table. They are used in relational databases to uniquely identify a given row, being analogous to the concept of the instance of an object in OOP. Just as two instances of the same object can be identical in terms of their attribute values, they are nevertheless distinct entities. Similarly, two rows in a relational database table can have the same values for every column but are still distinct records. Primary keys allow rows to be identified from each other regardless of the values of all other columns. Foreign keys are the primary keys of rows from other tables used to logically relate two rows from different tables together. Thus, a participant can be selected from the database and, using its primary key, all C3t instances that belong to it can be identified as long as the participant's primary key is the C3t instances foreign key.

Figure 21 shows the database structure as an entity relationship diagram.

Figure 21: Entity relationship diagram for the database backend. Each entity is a table in the database. PK is a primary key and FK is a foreign key. The relationship between the Participant and C3t table is a one-to-many-optional relationship, indicating a Participant row can have a relationship with zero or more C3t rows. The relationships between the C3t table and each task table is a one-to-one mandatory relationship, indicating every C3t row must have a single corresponding row in each Task table and vice-versa.

The data model of the database is significantly simpler to that of the C3t app, primarily because it only involves the relationships between tables, rows, and columns. Notably the concept of inherited entities is not included in the database model unlike the C3t model, as such abstractions are in direct contrast to the rigid structure relational databases rely upon. The main thing of note in Figure 21 is how the tables relate to each other. Participants have an optional one-to-many with C3t tests, indicating their primary key can act of a C3t rows foreign key between 0 and infinity times but a given C3t row can only record a single Participant primary key as their foreign key. The C3t table has a mandatory one-to-one relationship with the Task tables, indicating every C3t row must have a single row in each Task table whose foreign key is that C3t rows' primary key.

3.4.5 Use case implementation

3.4.5.1 Overview

Based on the now fully defined C3t and database data models, the use cases presented in section 3.3.2 can be more accurately written as follows.

1) CRUD operations for the Participant object/table
2) CRUD operations for the C3t and task objects/tables
3) Syncing a C3t instance with the PC used to configure the accelerometers

The following subsections will detail how each of these use cases is implemented in the RDCP. As the communication between the C3t app and server is the bedrock of the RDCP functionality, first an overview of how data is transferred between the C3t app and database backend is given. Afterwards, each of the remaining subsections will handle a specific use case showing the state of the system for each step of the use case flow alongside C3t app screenshots as appropriate. The high-level structure shown in Figure 18 will be modified showing the state of the system after each step. Descriptions are given in a step-by-step basis with possible points of errors or failures noted at each step. As has been previously stated, code will be kept to a minimum and will not be included due to its size in the appendix however the full source code can be made available upon request.

It should be noted at this point that one use case was not implemented – the ability to modify C3t information after it had been saved. The reason for this was that the C3t's developers felt the ability for sites and clinicians to modify test data at any point after the test was completed could potentially be harmful to the validity of collected data and that it was simpler to remove the functionality than heavily restrict its usage.

3.4.5.2 Transferring data between the C3t app and database backend

The communication between the C3t app and database backend shown in Figure 18 occurs over the internet. As noted in section 1.5, in order for such communication to be secure, data will need to be transferred over an HTTPS connection.

Communicating over an HTTPS connection generally involves sending one or more requests of various types of the connection. The types of requests, known as HTTP methods, are listed in Table 34.

Table 34: List of HTTPS methods and accompanying descriptions

HTTPS methods	Description
GET	Requests specified data from the recipient
POST	Sends a chunk of data to the recipient
PUT	Sends a chunk of data to the recipient but with the guarantee that two identical PUT requests will yield the same result
HEAD	Requests specified data from the recipient, but with only the meta data being returned not the whole requested resource
DELETE	Request to remove the resource from the recipient
PATCH	Request to make partial changes to an existing resource
OPTIONS	Request a description of available communication options from the recipient

Of these methods, GET, POST, DELETE and PATCH are used in the RDCP. The C3t app sends GET requests when it needs records from the database, POST requests when it wants to add records, PATCH requests when it wants to modify records, and DELETE requests when it wants to delete records. The reason PUT requests are not used simply because handling POST requests was simpler and, in terms of the C3t use cases, the functionality is identical. No use case of the C3t requires HEAD or OPTIONS messages to be sent.

Once the server receives a request, it process it then sends a response depending on the outcome of the request. The app is configured to handle the various response the server can send, displaying to the user the result of the attempted action.

For example, when saving a subject one of the things the database checks is whether the entered id code is already in the database. If it is not, and adding the participant succeeds, then a success code is sent to the app causing it to display the added participants details. If on the other hand the operation fails and the id code is already being in the database, then the server sends an error code which the app picks up on and informs the user the id code already exists.

The code for this operation is shown in Figure 22, where the *if* statements show the handling of the server response after a user has been added.

```java
if (jsonObject.getBoolean(ServerConstants.MEDIC_DB_KEY_SUCCEEDED)) {

    subjectToSave.setPrimaryKey(jsonObject.getLong(ServerConstants.MEDIC_DB_KEY_SUBJECT_PK));

    UtilityMethods.makeShortToast(getBaseContext(), "Participant saved successfully");

    Intent intent = new Intent();
    intent.putExtra(Constants.EXTRA_SUBJECT, subjectToSave);
    setResult(Constants.ResultCodeEnums.SUBJECT_CREATED.ordinal(), intent);

    finish();

} else if (jsonObject.getInt(ServerConstants.MEDIC_DB_KEY_ERROR_CODE) == 1) {

    basicSubjectDetailsFragment.displayDuplicateIDError();
    viewPager.setCurrentItem(0);

} else {
    throw new Exception(jsonObject.getString("Error Code: " + jsonObject.getString(ServerConstants.MEDIC_DB_KEY_ERROR_CODE)
            + "\nError Message: " + jsonObject.getString(ServerConstants.MEDIC_DB_KEY_ERROR_MESSAGE)));
}
```

Figure 22: C3t app code handling outcomes of adding a participant to the database

3.4.5.3 Use case 1: CRUD operations for the Participant object/table

3.4.5.3.1 Creating a participant

There are 7 steps to creating a participant using the RDCP as shown in Figure 23.

First, ‘create participant’ is selected from the home screen (Figure 23, bottom left) causing the create participant screen to be shown (Figure 23, second from left).

Second, the user enters participant details including study id code, date of birth, gender, and dominant hand. The id code can be restricted such that only valid study identifiers can be input (e.g., ‘HD-001’). By default, date of birth does not allow users under the age of 18 to be entered.

Third, once data is input the user selects ‘Done’ and the C3t app checks that all data has been correctly entered, displaying error messages as appropriate. If all data is correctly added the app asks for confirmation that the participant should be saved (Figure 23, bottom middle).

Fourth, if the user confirms the save the *json_encode* method mentioned previously for the Participant object packages the entered data and attempts to send a POST request to the database server. If an internet connection is not detected an error message is displayed to the user offering a link to the device’s internet settings. If the connection times out (indicating the server is down or otherwise unreachable) an error message is displayed to the user.

Fifth, assuming the data is sent and correctly received the server-side code checks the received data is valid for entry. If the data is valid the server attempts to save the received data in the database.

Sixth, the result of the save operation in the server-side code is transmitted back to the C3t app.

Seventh, the app receives the result of the save operation and displays it accordingly. There are three possible results; 1) the id code entered is already in the database, in which case this fact is displayed to the user asking them to change the id code, 2) the save for some reason failed in which case the user is prompted to contact the system admin, or 3) the save was successful in which case the participants profile screen is displayed (Figure 23, bottom right).

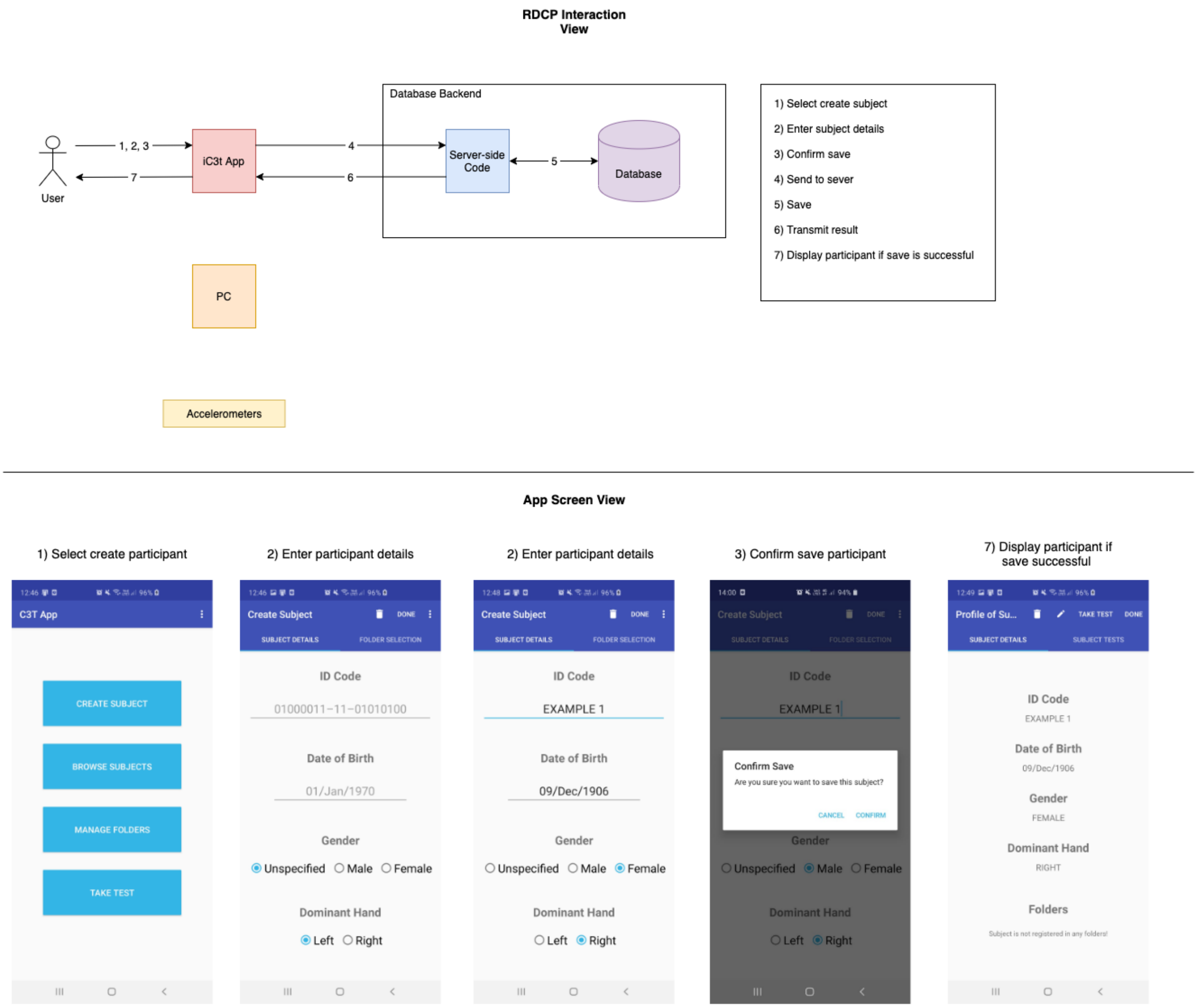


Figure 23: Steps for creating a new subject using the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.3.2 Viewing a participant (read)

There are 10 steps to view a specific participant's details as shown in Figure 24.

First, the browse button is clicked on the C3t app displaying an error if no internet connection is available (Figure 24, bottom left).

Second, if there is an internet connection a GET request is transmitted to the server for a list of all participants.

Third, the id codes and primary keys of all participants are extracted from the database.

Fourth, all id codes and primary keys are transmitted to the C3t app.

Fifth, a list of all participants is displayed to the user (Figure 24 middle).

Sixth, the user selects a specific participant.

Seventh, the primary key (PK) of the participant is packaged and transmitted in a GET request for that participant full details to the server.

Eighth, using the participants primary key, the participant details are extracted from the database along with the date of all C3t test instances they have stored.

Ninth, the data is packages into a json format and transmitted to the C3t app.

Tenth, the data is received on the C3t app, extracted, and used to generate the participant details screen (Figure 24, bottom right).

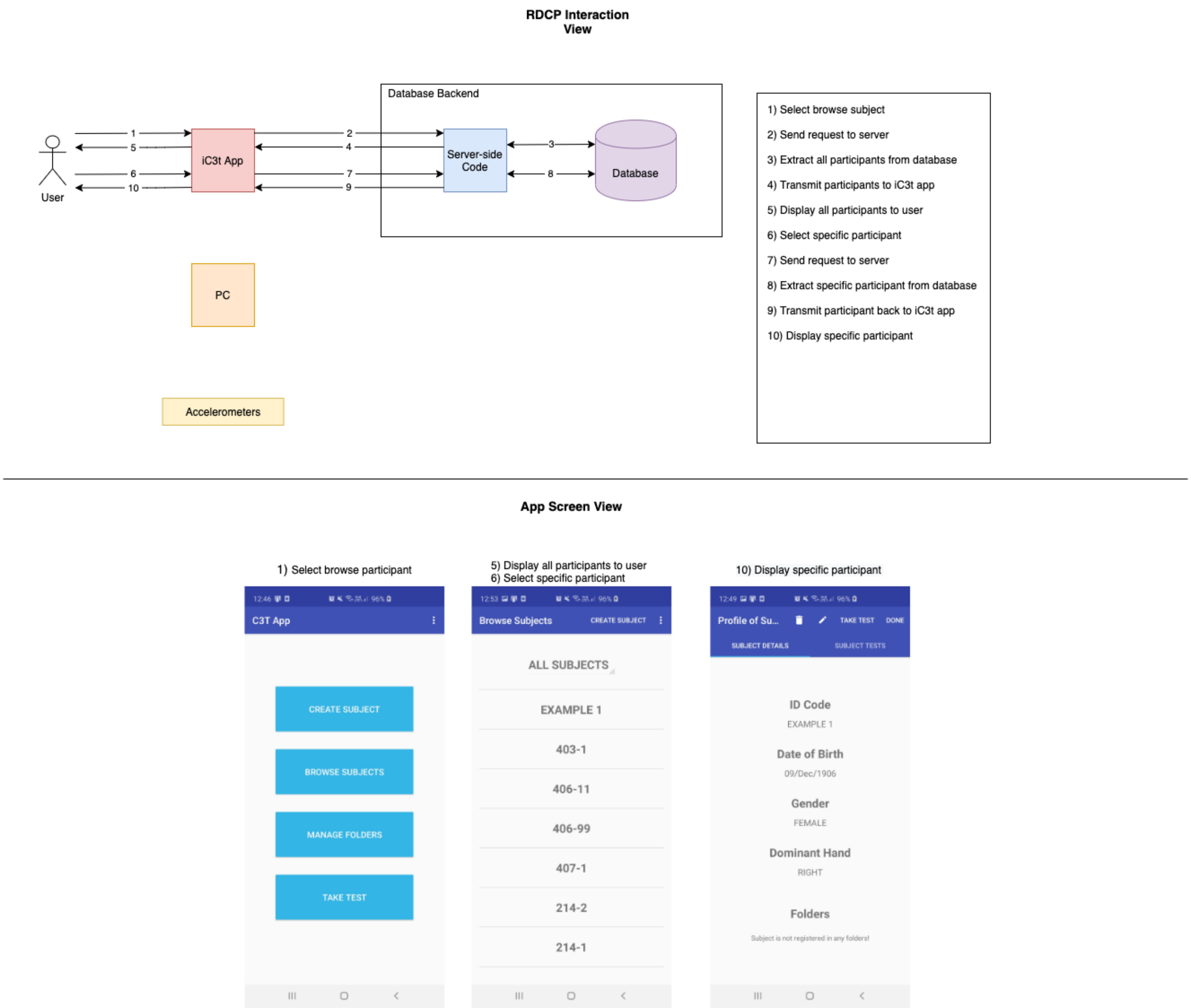


Figure 24: Steps for viewing a specific subject using the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.3.3 Updating a participant

There are eight steps to update a participant, starting from the participant profile screen, as shown in Figure 25.

First, the pencil icon at the top of the participant profile screen is selected (Figure 25, bottom left).

Second, the same screen used to create a participant is shown but populated with the participant details (Figure 25, bottom, second from the left).

Third, any alterations to the participant details are made.

The fourth, to eighth steps are functionally the same as third to seventh steps in Creating a participant – the user confirms the update, causing the update to be transmitted to the server which attempts to alter the participant in the database before transmitting the results of the operation. The only two differences are that a PATCH request is sent to the server rather than a POST request, and if the update fails on the server side, the original participant details will be retained unaltered in the database.

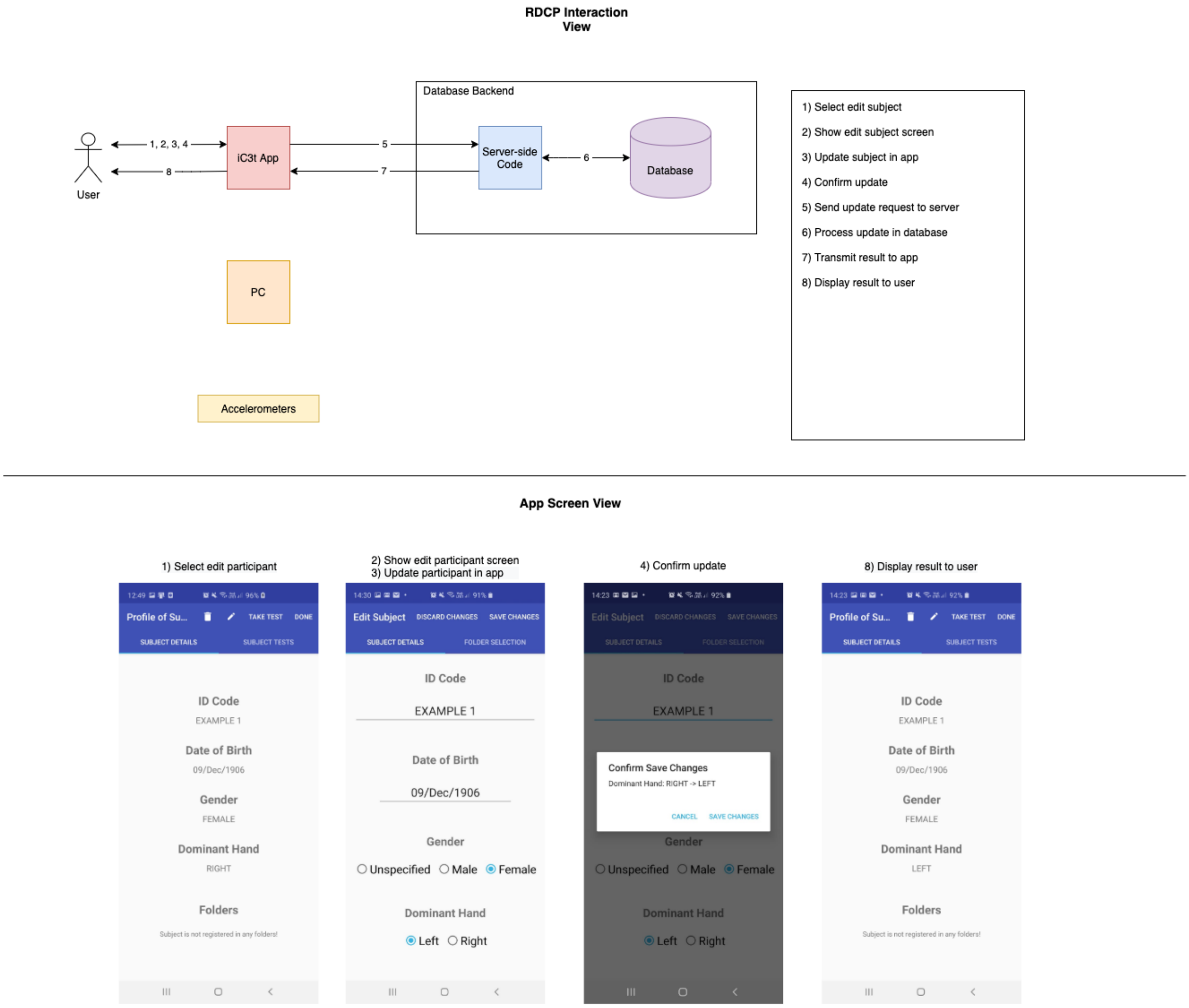


Figure 25: Steps for updating a participant using the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.3.4 Deleting a participant

There are eight steps to delete a participant from the database, starting from that participants profile screen, as shown in Figure 26.

First, the delete participant button, the dustbin icon at the top of the screen, is selected by the user (Figure 26, bottom left).

Second, the user is asked to confirm the deletion (Figure 26, bottom middle)

Third, upon deletion confirmation, the primary key of the participant is sent in a DELETE request to the server, with an error being displayed if no internet connection is detected.

Fourth, the DELETE request is received by the server and processed, attempting to delete the participant and all associated C3t records in the database.

Fifth, the result of the DELETE request is transmitted to the C3t app.

Sixth, the result of the DELETE request is received by the app and the result displayed accordingly. If the deletion was unsuccessful this will be displayed in a popup message on the participants profile screen. Otherwise, the browse participant screen will be displayed with the deleted participant removed (Figure 26, bottom right).

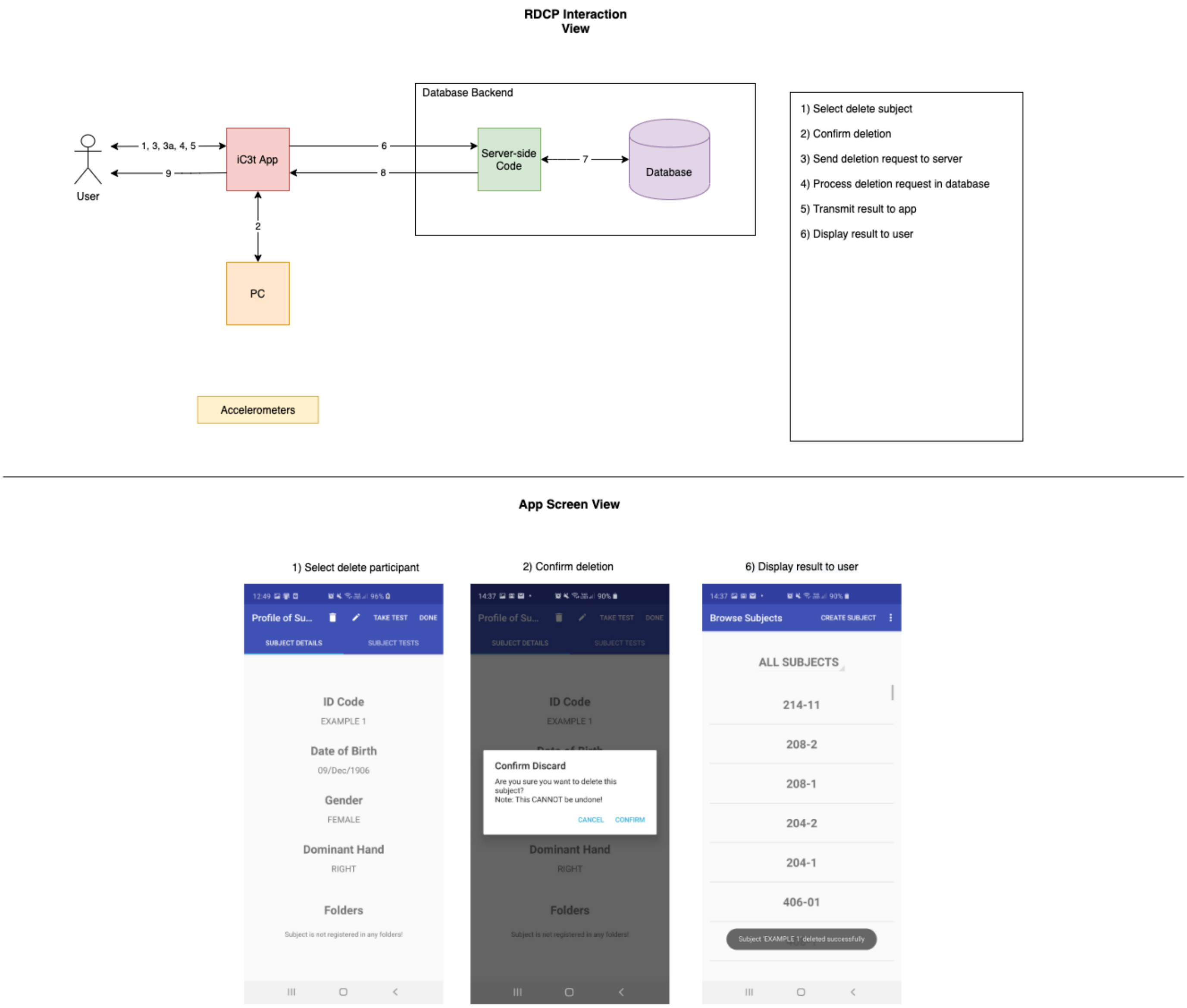


Figure 26: Steps for deleting a participant from the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.4 Use case 2: CRUD operations for the C3t and task objects/tables

3.4.5.4.1 Taking the C3t (create)

Taking the C3t using the C3t app is a fairly involved process. Broadly, the user selects a participant, selects the 'take test' button, synchronises the app with the accelerometers if they are in use, and then proceeds through each of the task screens taking each in turn. As shown in Figure 27, there are 9 main steps with 3 sub-steps per C3t task, starting from a participant's profile screen.

First, the take test button is selected (Figure 27, upper row, left).

Second, the popup used to synchronise the C3t app and accelerometers is shown (Figure 27, upper row, middle). The synchronisation process is detailed later in section 3.4.5.5.

Third, the user is displayed with button for each of the 6 C3t tasks (Figure 27, upper row, right). Each task must be taken in the order specified in the C3t manual (BTT, BVT, CVT, BAT, CTT, DTT) and cannot be taken out of order. If a user tries to take the wrong task next an error is displayed telling them which task to start. Uncompleted tasks are shown in blue, completed tasks are shown in blue.

Each task has a dedicated screen which contain the functionality to record that tasks results. All task rules listed in the C3t manual are encoded into the app. Task instructions are also encoded in the app and can be viewed on each tasks screen by selecting the. Once a task is completed the user has the option to restart and view a task. Viewing or saving a task will cause all derived variables to be calculated and stored.

Fourth, after all tasks have been taken is to save the test. Saving a test is not possible until all tasks have been taken. If no internet connection is available, the usual popup appears informing the user and linking them to their device's internet settings.

Fifth, whether the user is asked to confirm they would like to save the test.

Sixth, the results of the test are packaged using the *json_encode* method for each C3t and task object before being sent via a POST request to the server.

Seventh, the test is unpackaged by the server and recorded in the database.

Eighth, the results of the save operation are placed into a response and transmitted to the C3t app.

Ninth, the results of the save operation are displayed to the user. Whether or not the save was successful, full test results are shown to the user. If the save was unsuccessful, the user is informed and asked whether they would like to retry the save or complete a paper CRF.

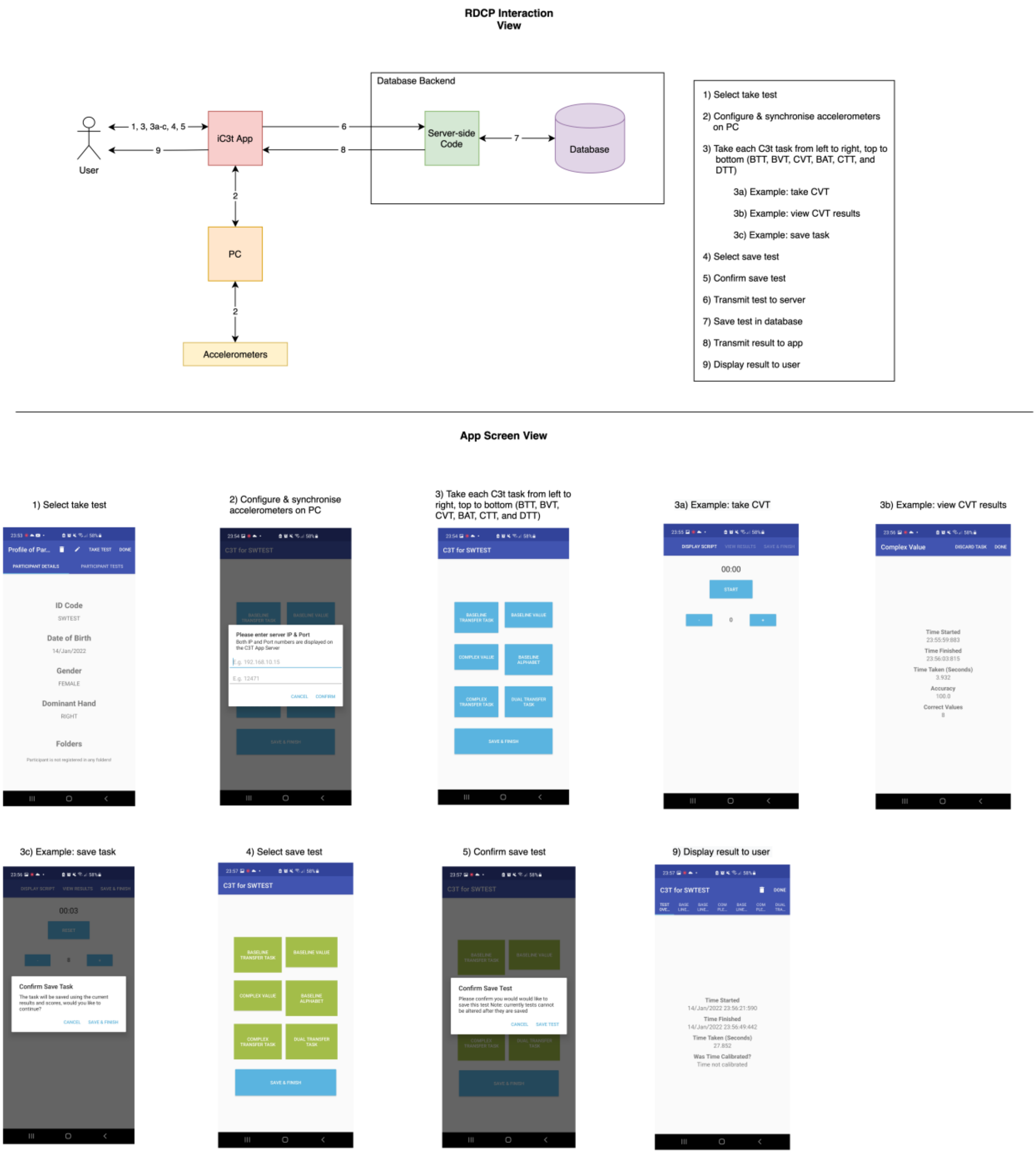


Figure 27: Steps for taking the C3t using the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.4.2 Viewing a C3t instance (read)

There are six steps to view a C3t instance, starting from a participant's profile screen, as shown in Figure 28.

First, the user navigates to the test list on a participant's profile (Figure 28, bottom left).

Second, the user selects a test from list (which are displayed using the time and date they were taken).

Third, a GET request is sent to the server using the tests primary key.

Fourth, the test is extracted from the database using its primary key.

Fifth, the test is packaged into JSON format and sent back to the C3t app.

Sixth, the test data is extracted by the app and used to display the test results to the user (Figure 28, bottom right).

Figure 28: Steps viewing a C3t instance stored in the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.4.3 Deleting a C3t instance

There are seven steps to deleting a C3t instance, starting from a test view screen, as shown in Figure 29.

First, the delete test button, the dustbin icon at the top of the screen, is selected by the user (Figure 29, bottom left).

Second, the user is asked to confirm the deletion (Figure 29, bottom middle)

Third, upon deletion confirmation, the primary key of the test is transmitted in a DELETE request to the server, with an error being displayed if no internet connection is detected.

Fourth, the DELETE request is received by the server and processed, attempting to delete the C3t instance and associated tasks using the C3t instance's primary key.

Fifth, the result of the DELETE request is transmitted to the C3t app.

Sixth, the result of the DELETE request is received by the app and the result displayed accordingly. If the deletion was unsuccessful this will be displayed in a popup message. Otherwise, the participants profile screen will be shown with the deleted instance removed from the test list (Figure 29, bottom right).

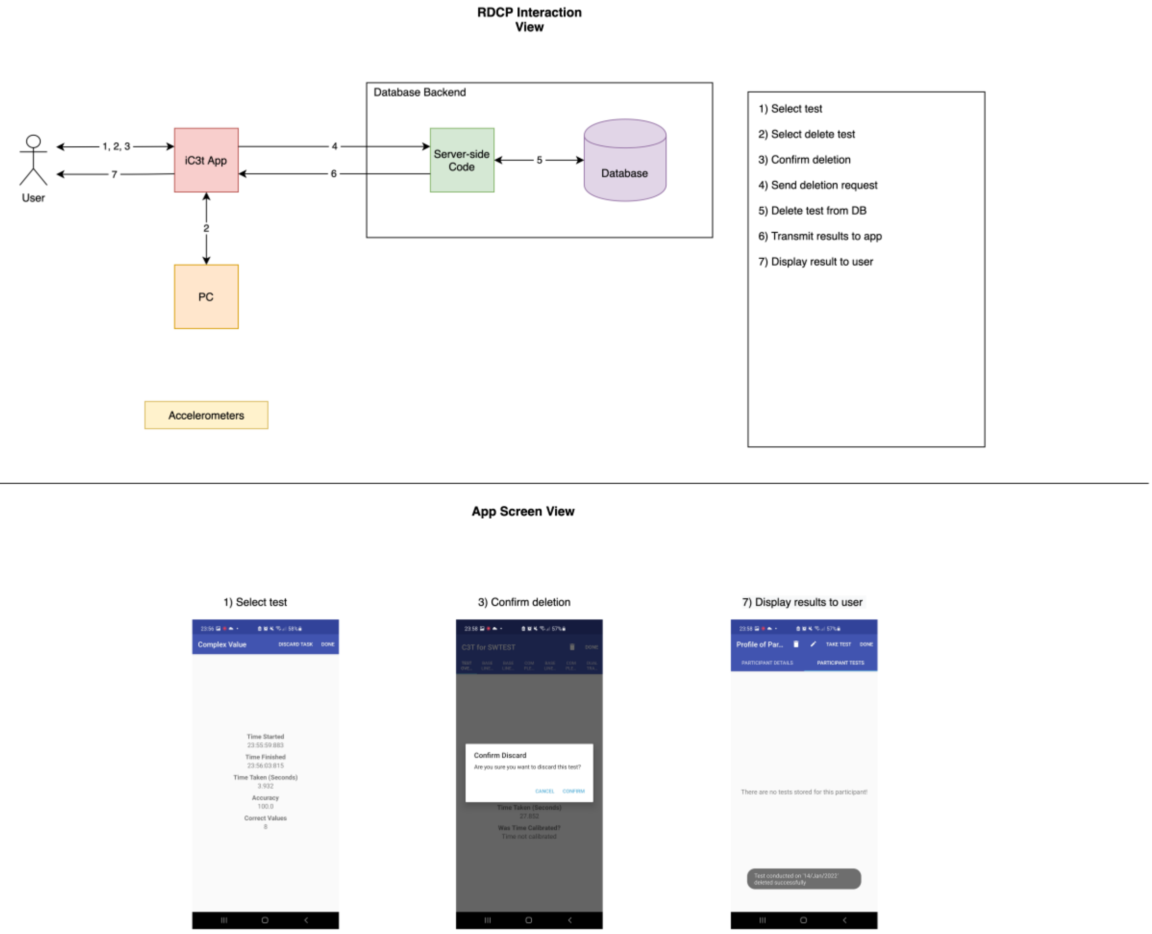


Figure 29: Steps for deleting a C3t instance stored in the RDCP. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The

screenshots at the bottom of the figure show the various screen the C3t app goes through with the relevant step noted at the top of each image.

3.4.5.5 Use case 3: Syncing a C3t instance with the PC used to configure the accelerometers

An integral part of the C3t app is the ability to synchronise timestamps between the C3t app and the computer used to configure the accelerometers. It should be noted once again that the RDCP does not facilitate the collection of accelerometer sensor data but does allow for interoperability between the C3t app and accelerometers. There are six steps to synchronise the timestamps, starting from a participant's profile screen, as shown in Figure 30.

First, the accelerometers need to be configured on an internet connected computer using GeneActiv's software. This will set their internal clocks to the clock of the computer.

Second, the C3t PC app should be opened on the computer used to configure the sensors (Figure 30, bottom left). The C3t PC app will start a server on the PC, displaying its IP address and port number then wait for a connection from the app.

Third, in the C3t app, the user should select to take a new test.

Fourth, in the popup screen shown on the C3t app, the user should enter the IP address and port number shown on the C3t PC app into the appropriate fields and then press confirm (Figure 30, bottom, second from the left).

Fifth, the C3t app will send a connection to the C3t PC app. For this step, the C3t app and PC app must be connected to the same network. If the synchronisation is successful a message will be displayed on the C3t PC app (Figure 30, bottom, third from the left). The C3t app will also display a popup message and allow the C3t to be taken (Figure 30, bottom right). Finally, once synchronisation is complete the C3t can be taken.

If the synchronisation was not successful, the user will be informed and given the option to retry the synchronisation or continue without the synchronisation. If synchronisation is not performed a field will be set in the C3t instance and stored in the database indicating that it was not performed and as such the C3t instance and accelerometer timestamps may be out of sync with each other.

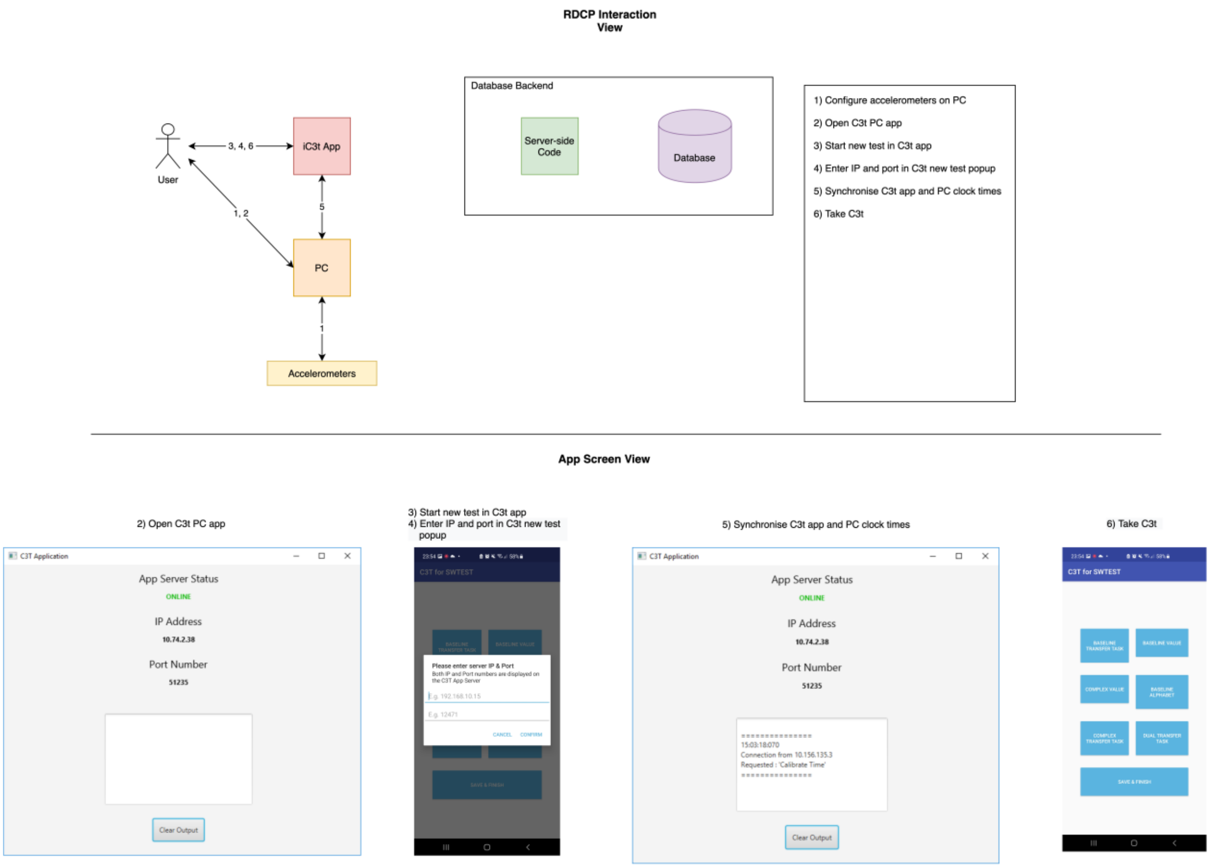


Figure 30: Steps for synchronising the instrumented C3t and C3t PC app. The top left-hand corner shows the interaction between the various RDCP components. The top right-hand corner lists the steps in order of execution. The screenshots at the bottom of the figure show the various screen the C3t app and C3t PC app goes through with the relevant step noted at the top of each image.

The synchronisation process itself is performed using the Simple Network Time Protocol (SNTP) originally defined in RFC 4330 (Mills, 2006). The SNTP is a simplified version of the Network Time Protocol and facilitates the synchronisation of clock times between two computers over a network. The steps of the SNTP are as follows.

1) The C3t app gets its current system time (T1) and sends it to the C3t PC app
2) The C3t PC app records its system time T1 was received (T2)
3) The C3t PC app gets its current system time again (T3) and transmits it, along with T2, back to the C3t app
4) The C3t app records the system time (T4) that it receives T2 and T3
5) The C3t app uses T1, T2, T3, and T4 to calculate the delay across the network and the difference between the clock times

An example of this process is shown in Figure 31.

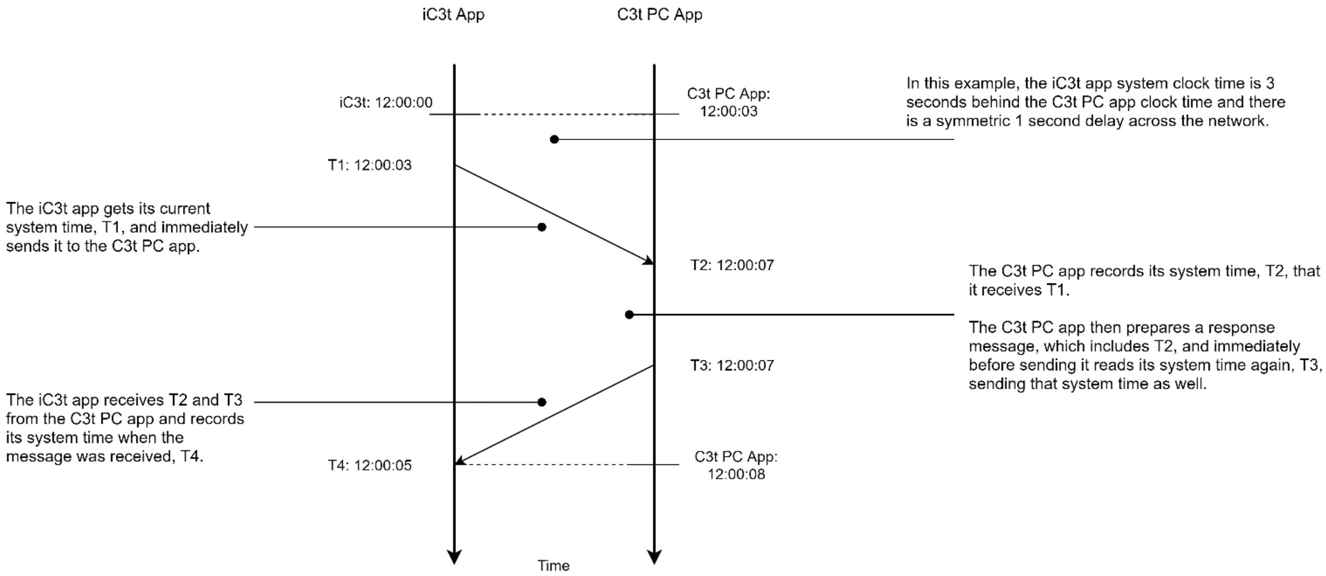


Figure 31: SNTP steps. By sending timestamps between the C3t app and C3t PC app the difference between the two system clocks can be calculated.

Using the following equations, we can calculate the difference between the C3t app clock time and the C3t PC app clock time (the clock offset).

$$T4 = T3 + roundtrip\ delay + clock\ offset$$

So;

$$clock\ offset = T4 - T3 - roundtrip\ delay$$

Where;

$$Roundtrip\ delay = \frac{(T4 - T1) - (T3 - T2)}{2}$$

For the example in Figure 31 (where we know the offset is 3 seconds and there is a 1 second delay across the network) the above equations would produce the following results.

$$Roundtrip\ delay = \frac{(12{:}00{:}05 - 12{:}00{:}03) - (12{:}00{:}07 - 12{:}00{:}07)}{2} = \frac{2 - 0}{2} = 1$$

$$Clock\ offset = 12{:}00{:}05 - 12{:}00{:}07 - 1 = 5 - 7 - 1 = -3$$

Thus, the C3t app system clock is 3 seconds behind the C3t PC app and so the accelerometers. Even a difference as small as this would have a dramatic impact on the reliability of the sensor readings. Assuming the recording frequency of the accelerometers is 100Hz, the C3t app clock time being 3 seconds behind would result in 300 samples being added before the task was started and 300 samples at the end of the task being missed.

3.5 RDCP Evaluation

3.5.1 Overview

Section 3.3 and 3.4 detailed the design requirements and implementation of the RDCP respectively. This section evaluates how well the RDCP performed by assessing its performance across the three separate studies it was embedded into over the course of this project (see section 3.5.2.1.). The performance of the RDCP is critically evaluated based on its observed performance across three studies it was embedded into. Based on the RDCPs performance, recommendations are then made for future similar projects. For the purpose of evaluating the RDCP, the accelerometers used in each of the studies the RDCP was embedded into (GeneActiv tri-axis accelerometers, ActivInsights; UK) are also evaluated. This is because the RDCP facilitates the collection of and was designed to work with these sensors. As will be shown, the RDCPs performance was somewhat damaged by the inclusion of these particular sensors. Additionally, as has been mentioned several times throughout this chapter and in chapter 1, the waterfall methodology was used for the construction of the RDCP and so will be briefly discussed.

3.5.2 Performance Evaluation

3.5.2.1 Embedded Studies

The RDCP was embedded into three of the studies listed in chapter 2 – PACE-HD, TRIDENT and Developing Clinical Applications for a Novel Multi-Task Functional Assessment: The Clinch Token Transfer Test (referred to here as C3t PhD). For information on each of the studies see section 2.3.1.1. It should be noted that C3t PhD and TRIDENT were both single-visit studies whilst PACE-HD had data collected twice, once at a baseline visit and once at a 6-month follow-up visit.

At the request of the study managers for PACE-HD some functionality of the C3t app was disabled. Specifically, at the request of the study's primary investigator, sites were only allowed to create a participant and then immediately take a C3t instance. If sites created a participant and then exited the app there was therefore no way for them to re-access the participant. The rationale for this was that sites should not have the ability to alter test details themselves after tests had been completed. This, in theory, should have been fine, however as is shown in the following sections did cause some errors to occur that would otherwise have been avoided.

3.5.2.2 Data Collection Evaluation

The purpose of the RDCP is to facilitate the collection of C3t and accelerometer data from study sites and store that data in a remote database hosted at Cardiff University. As such, the primary metric that can be used to judge its suitability is how much of the data that could have been collected for each study was correctly collected and ultimately usable.

The data the RDCP facilitates the collection of can be split into two categories – C3t data, which is directly collected using the C3t app, and sensor data the collection of which is facilitated the RDCP (via the time synchronisation described in section 3.4).

Please note that the RDCP facilitates the collection of sensor data but does not handle its transfer. As described in section 3.3.2, the RDCP handles the synchronisation functionality between the C3t app and GeneActiv sensors. In two of the three studies, C3t PhD and TRIDENT, sensor data was collected and analysed at the same site (Cardiff University, UK) and so no transfer was necessary. PACE-HD however was a multi-centre study and so FastFile, an online software platform for sending and receiving files, was used to transfer sensor data for this study. All C3t data was collected and transferred via the C3t app for all studies.

Usable data was defined differently for the C3t and sensor data. C3t data was considered to be complete and usable if all C3t variables were stored for a given participant in the server database. Sensor data was considered to be complete if all files (dominant hand sensor and non-dominant hand sensor) were sent via FastFile. Sensor data was considered usable if the timestamps of the relevant C3t tasks could be found within the file and if the accelerometer recording passed visual inspection (i.e., upon observing the data it looked to contain a valid C3t recording). Ideal and corrupt C3t recordings are shown in Figure 32 and Figure 33 respectively. The timestamps of the BTT and CTT tasks only were required to be found in the sensor data files. This is because (as is discussed in chapter 4) an additional set of BTT and CTT task data was available for analysis from a previous version of the C3t, but that previous version did not contain the final transfer task (the DTT).

Table 35 shows the maximum number of C3t records that could have been available (i.e., complete & usable) and the total number that were available across the studies and overall. Table 36 shows the maximum number of sensor records that could have been available (i.e., complete & usable), the number of complete records, the number with correct timestamps, and the number that were ultimately deemed usable.

Table 35: Collection results for C3t data. The maximum possible column indicates the total number of samples that should have been collected during each study. The number collected column indicates the number of samples that were ultimately found to be complete in the database. Completeness is defined as all C3t task information being present in the database. The matched baseline & follow-up row refers to paired C3t instances collected for participants across baseline and follow-up (i.e., matched records across baseline and follow-up collections).

Study Performance			
Study	Maximum Possible	Number Collected	% of maximum
PhD	20	20	100%
Trident	20	17	86%
PACE (baseline)	60	52	86%
PACE (follow-up)	60	39	65%
PACE (matched baseline & follow-up)	60	37	61%
Total usable samples (baseline): 89 of 100 (89%) **Total usable samples (follow-up): 39 of 60 (65%)** **Total usable sample (batched baseline & follow-up): 37 of 60 (61%)**			
PACE site performance			
Site	Maximum Possible	Number Collected	% of maximum
PACE Site 1 (baseline)	20	23	115%
PACE Site 2 (baseline)	20	12	60%
PACE Site 3 (baseline)	20	17	85%
PACE Site 1 (follow-up)	20	17	85%
PACE Site 2 (follow-up)	20	10	50%
PACE Site 3 (follow-up)	20	12	60%

Table 36: Collection results for sensor data. The maximum possible column indicates the total number of samples that should have been collected during each study, taking into account the 'number collected' column from Table 35. The complete records column indicates the number of samples for which all files were present. The correct timestamps column indicates the number of samples for which all files were present and the BTT and CTT task timestamps could be found in. The usable records column indicates the number of samples which upon visual inspection of the BTT and CTT task timestamps appeared to contain valid data (e.g., not a flatline indicating no collection). The matched baseline & follow-up row refers to paired C3t instances collected for participants across baseline and follow-up (i.e., matched records across baseline and follow-up collections).

Study Performance				
Study	Maximum Possible	Complete Records (all files present)	Correct timestamps (able to find C3t timestamps in all files)	Usable records (after data inspection)
PhD	20	17	17	17
Trident	17	16	15	15
PACE (baseline)	52	40	26	22
PACE (follow-up)	39	8	2	1
PACE (matched baseline & follow-up)	37	0	N/A	N/A
Total usable samples (baseline) 54 of 89 (60%) **Total usable samples (follow-up) 1 of 39 (2%)** **Total usable samples (matched baseline & follow-up) 0 of 37 (0%)**				
PACE site performance				
Site	Maximum Possible	Complete Records (all files present)	Correct timestamps (able to find C3t timestamps in all files)	Usable records (after data inspection)
PACE Site 1 (baseline)	23	20	14	11
PACE Site 2 (baseline)	12	8	0	0
PACE Site 3 (baseline)	17	12	12	11
PACE Site 1 (follow-up)	17	1	1	0
PACE Site 2 (follow-up)	10	1	0	0
PACE Site 3 (follow-up)	12	1	1	1

Figure 32: An ideal C3t accelerometer recording (raw data), note the 8 distinct overall peaks & troughs which likely correspond to the 8 tokens in the assessment and the range of values of G.

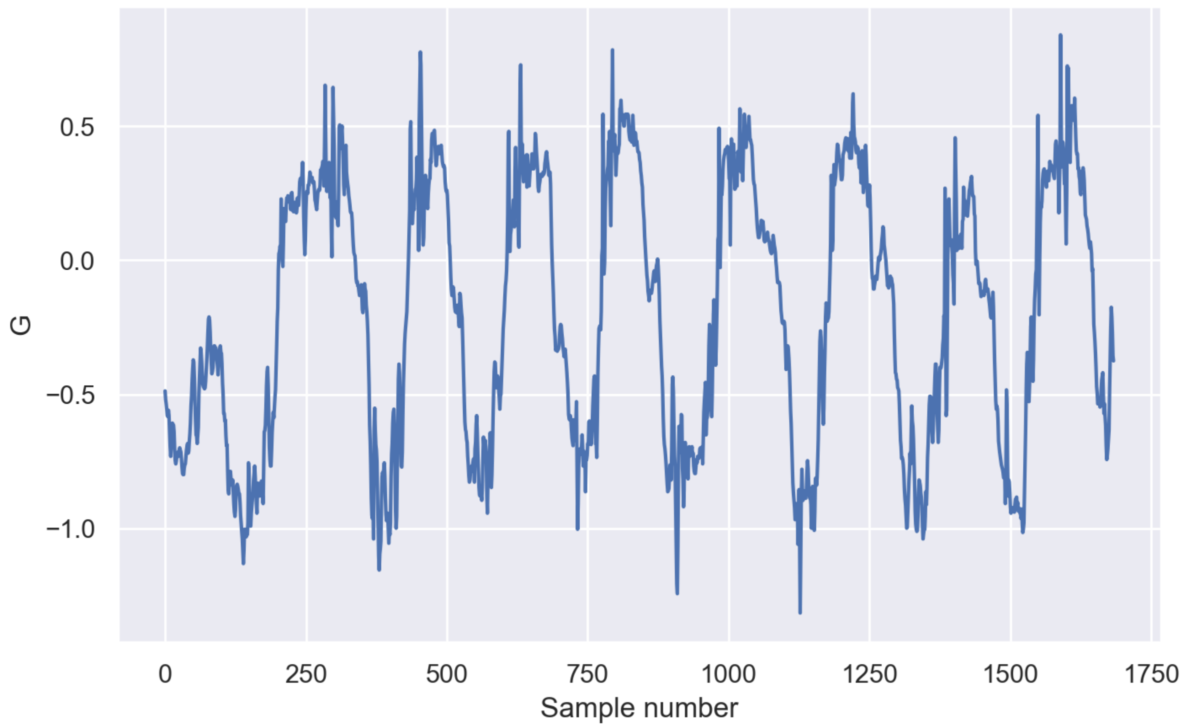
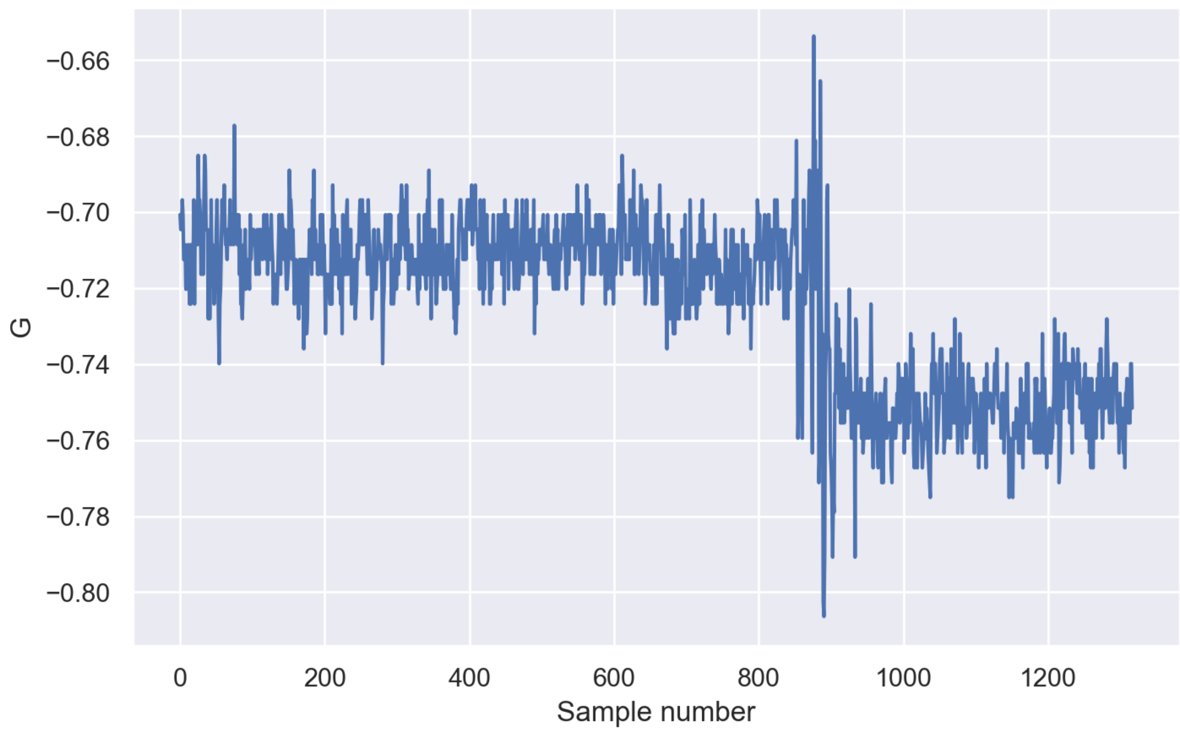


Figure 33: A corrupt C3t recording (raw data), note the lack of any discernible structure and the small variation in G.



Based on the tables above, the RDCP appeared to perform well for the C3t data but poorly for the sensor data. The following summarises the results shown in Table 35 and Table 36, sections 3.5.2.3 and 3.5.2.4 explore why poor performance was observed in some cases.

Table 35 shows 89% of the C3t data that could have been collected at baseline across all 3 studies was collected correctly. All data from C3t PhD was correctly collected, TRIDENT is missing 3 samples and PACE is missing 8 samples. There was a significant drop in collection quality in the PACE follow-up dataset from 86% at baseline to 65% at follow-up.

Table 36 shows only 60% of the sensor data that could have been collected was ultimately collected and was usable. In C3t PhD 3 samples were incomplete but all complete samples were usable. In TRIDENT 1 sample was incomplete and 1 sample did not contain the correct timestamps for use. In PACE 12 samples were incomplete, 14 samples did not contain the correct timestamps and 4 samples were declared unusable after visual inspection. Similar to the C3t data, there was a significant drop off in sensor data quality at the follow-up visit, to the point that only 1 usable file was obtained.

In both Table 35 and Table 36 there is clear difference between data collection quality across the sites. Site 2 performed significantly worse than sites 1 & 3. At baseline, 63% of its C3t data was properly collected and not one sensor file contained the correct timestamps.

3.5.2.3 User experience evaluation

Based on Table 35 and Table 36, there were clearly some problems with the RDCP it is important to consider why these problems may have occurred. Table 37 shows a list of all emails received from across all studies regarding errors and queries to do with using the RDCP for data collection. Each assigned a problem type as follows.

- A ‘user error’ indicates some action or inaction on the user’s part caused the problem
- A ‘technical fault’ indicates that the problem was either a software bug or could have been avoided with additional software development

Table 37: Emails received from sites during the studies with issues regarding some aspect of the RDCP. Columns include the date the email was received, the description of the problem, the resolution was applied and a problem type. Problem types are 'user error', 'technical fault', 'user error / technical fault', or 'N/A'. A user error is defined as a problem that occurred due to some action or inaction that the user took. A technical fault is defined as a problem that occurred due to a software bug or otherwise could have been avoided programmatically. One problem was never resolved or fully discussed and thus is listed as N/A.

Date of first email	Problem Description	Resolution	Problem Type
06/08/2018	Site could not synchronise C3t app and C3t PC app	Site was trying to enter example IP and port connection details from manual and had not installed the C3t PC app, after installation problem was resolved	User error
24/10/2018	Site complained that their hand calculations of the derived variables did not match those reported by the app	No resolution, site did not respond to subsequent emails	N/A (unclear if it was a problem or a miscalculation)
17/11/2018	Site could not synchronise C3t app and C3t PC app	Site was not connected to the internet, after connecting the problem was resolved	User error
25/01/2019	Site reported an issue with the keyboard not displaying the dash symbol (-) which is needed to properly enter PACE participants	Site had inadvertently changed the keyboard type in use to a non-standard one, a patch was created to stop them from doing this in the future	Technical fault / user error
29/01/2019	Site (different from 25/01/2019) reported bug with keyboard not displaying the dash symbol again (-)	New android version had been released which altered the layout of the used keyboard, a patch	Technical fault

		was created to give sites access to the full keyboard but limit the characters they could enter	
11/02/2019	Site created a participant then exited the app, leaving them unable to access that participant again	Manually removed participant from the database allowing them to enter that participant's details again	Technical fault / user error
18/02/2019	Site reported the same error as on 29/01/2019	Site had not installed the distributed app patch	User error
23/03/2019	Site could not synchronise C3t app and C3t PC app	Site was not connected to the internet, after connecting the problem was resolved	User error
09/04/2019	Site was unable to create a participant in the database as the ID was already in use	Site had already created this participant previously whilst testing the app, as they could not access previously stored participants the erroneous participants had to be removed from the database manually	Technical fault / user error
24/04/2019	Site broke the tablet and had bought a new one to replace it however as the new tablet used a newer version of Android the app did not work for it	Updated app to be compatible with the newer version of Android	Technical fault / user error
17/05/2019	Site could not synchronise C3t app and C3t PC app	Site did not have the internet turned on	User error
28/11/2019	Site reported one of the GeneActiv devices was not collecting data	Site had not properly connected the device to	User error

		the charger causing it to run out of battery before the next time it was used and so no data was collected	

Table 37 shows there were 12 query emails sent, 6 of which were user errors, 1 of which was a purely technical fault, 4 of which could be considered either a technical fault or a user error and 1 of which is unclassified as it is unclear whether there was a problem.

The query emails suggest that there were issues with both the app itself and with its use by the users. Four separate emails were received where users could not properly synchronise the instrumented C3t and C3t PC app because they did not have their internet connection turned on. This is surprising as the C3t app displays an error message when an internet signal is not detected.

Some of the errors were technically caused by the users however with more robust programming and design decisions some of these issues may have been avoided. For example, the app should have been kept up to date with the latest android releases, however it was assumed that sites would only be using the tablets provided which had a version of android the C3t app is compatible with. Two sites reported issues with the keyboard layout which could have been avoided had the software prevented them from inadvertently changing the keyboard layout. Similarly, two sites were unable to collect data as they could not access participant profiles, they had setup in advance, which could have been avoided had that functionality not been removed from the app for PACE.

Other errors included sites not installing software patches, not charging the sensor before collection, and not properly following the manual causing them to input the example IP and port details shown in the manual instead of the actual ones displayed by the C3t PC app.

3.5.2.4 RDCP Performance Discussion

The RDCP performed well in terms of collecting C3t data, however issues were clearly encountered for the sensor data and there is room for improvement for both use cases.

The C3t data is in theory very simple to collect with the app. However, it does make certain assumptions about usage. One of these assumptions that seemed to prove problematic was the presence of an internet connection. There is no way to transfer data without an internet connection and whilst recorded data could have been stored on the app and uploaded later this comes with its own complexities. For example, if two sites create new participants with the same ID code and neither uploaded data to the server both could attempt to uploaded it at a later date causing conflicting data to be added. It also complicates data management, as data can potentially be stored at a site on a mobile device without the data managers necessarily being aware of where the device is being stored when not in use (i.e., is it locked away properly).

Similarly, the app was not originally designed to restrict functionality. The inability to recall participants from the database caused problems for some of the sites. This could have been avoided programmatically, for example if all participants are essentially 'one shot' entries into the database then there is no need for participant creation and taking the C3t to be separate. Unfortunately, this was not in the original design specification of the app and thus was not accounted for.

Regarding the GeneActiv accelerometers, the lack of emails asking for assistance suggests that the sites thought the devices were being used correctly. However, clearly there were problems that prevented sites from collecting high-quality sensor data. Numerous files either did not contain any data (with some files containing data recorded days or even weeks before the C3t instance was taken) or were not present at all. Notably some files were extremely large (exceeding 10GB) and yet contained no relevant data, suggesting they had accidentally been switched on to record and never configured properly again. These problems, coupled with the lack of emails asking for assistance, suggest that the issue encountered with the GeneActiv accelerometers was that sites thought they were collecting data correctly but were in fact not. It is plausible that the reason for this is the lack of feedback the sensors provide to confirm they are currently recording, and this will be discussed in the next section. Overall, it can be concluded that (at least from the perspective of this study) GeneActiv accelerometers are ill-suited for clinical data collection. Overall, given the difficulties encountered using GeneActiv accelerometers during this study, alternatives may be preferred for similar multi-centre studies where technicians cannot always be present when data are being collected.

3.5.2.5 Waterfall Development Methodology

The Waterfall methodology of development (see section 1.5.3.2.2.3) was used in this study to construct the RDCP. Whilst not technically part of the performance of the RDCP the impact this methodology had on the general development lifecycle is important to consider and critique.

The rationale for using Waterfall, which is typically seen as a slightly outdated development methodology (again see section 1.5.3.2.2.3) in this project was that the more iterative approach offered by Agile was unsuitable due project deadlines. Additionally, due to the perceived stability of the RDCP's requirements the main drawback of Waterfall (that it is brittle under requirement changes) was not relevant.

However, during the RDCPs development the MBT was updated to its most recent version, the C3t, and so the RDCP had to be updated to account for the changes to the test. At the time the update was required the RDCP was by-large developed and the various components designed & implemented. To account for the test update large portions of the system had to be updated and redesigned. Whilst the RDCP itself was not negatively impacted the overall development time increased as a result, resulting in less time for other aspects of this project (e.g., data collection & analysis). This is a classic case of act of the when the Waterfall methodology falls down – it is completely unsuitable when requirements change during a project which even in research projects which should be highly regulated timewise can occur.

3.5.3 Recommendations for future studies

Systems like the developed RDCP are becoming ever-more common in medical research and, as has been discussed, often form the backbone of modern-day research. As such, critically analysing and learning from previous development efforts of such systems is important for streamlining future development efforts. To this end, based on this study there are a number of recommendations that can be made for future systems built to facilitate remote data collection of instrumented assessments in a multi-centre study.

First and foremost, different sensors should be used in future projects. The rationale for this is that the issues encountered were data quality issues which can only be attributed to the sensors design (as the quality of the sensor data depends on nothing else). By default, recording occurs once the sensors are removed from their charging cradle. However, there is no indication that the sensors are recording or how much battery life they have left (unless they are plugged into the cradle). This is likely what resulted in much of the PACE sensor data to either not contain the right timestamps or to just not be collected at all. This recommendation is in line with the recognised ongoing problem of picking the correct sensor for research studies, particularly in fields which are newer to working with such devices such as medicine (Russell *et al.*, 2021). A broad recommendation is that future work in fields which require the usage of sensor technology should consider implementing a proper review & feasibility study process prior to deciding which sensor set to use, as is suggested by Rosa *et al.*, (2021). It should also be noted however that following a so called 'adaptive study design', which incorporates flexibility of data collection into the underlying study design is an alternative which can be considered (Pallmann *et al.*, 2018). An adaptive study design has the benefit of allowing data collection to begin sooner than if a rigorous review process is carried out, whilst allowing for changes in sensor selection during the study if required. Overall, future work should pay close attention to the choice of sensor selection, and either conduct a rigorous feasibility study prior to beginning data collection or make allowances in the study design for direction changes if required.

Second, future projects which seek to develop software similar to the C3t app should focus on understanding all potential projects the software will likely be integrated into and what requirements these may have. The difficulty of having different requirements per project is that even seemingly small changes to functionality can require dramatic changes to a codebase. For example, it was a requirement of PACE that sites not be able to access the remote database, only upload data to it. However, altering the app such that participants would not be saved until a C3t instance had been taken would have required a significant re-write and so the decision was made to simply remove the functionality and trust that the sites would use the software as intended. This however proved to be naïve and as a result some data was lost. These sorts of specialised requirements are part of the reason that at present there is such an appetite for custom software solutions, particularly within clinical trial research. Whilst there are a multitude of 'plug and play' commercial options available, a recent review found that many trial designs were unsupported (Meyer *et al.*, 2021). This will naturally be exacerbated in cases like the one presented here, where unique specific combination of technology and assessment (i.e., accelerometers & the C3t) are unlikely to be covered by commercial options. Ultimately, a proper requirements gathering phase could have avoided some of the issues encountered by this study, and future work is advised to incorporate a requirements gathering phase, as is also suggested in the literature (Inan *et al.*, 2020; Rosa *et al.*, 2021).

Third, whilst the Waterfall methodology of development was felt to be appropriate for this study the problem it has with changing requirements was still encountered. The timelines of research projects tend to be quite well defined as such timelines often constitute part of funding applications. As such, in theory, software required for such research projects should have their requirements fully defined before the project, and so development, begins. Thus, it makes sense that Waterfall, a development methodology which focuses on moving along specific stages of development would be preferable to Agile which focuses on flexibility between those stages. However, in this project it was found that even though the specifications were written and thought to be finalised the requirements of the software changed nonetheless and so exposed the projects to Waterfall's brittleness. This issue of choosing an ideal software development methodology to use in medical research is not new, and has led to alternative methods being developed which modify Agile to fit into a more rigid structured environment (Özcan-Top and McCaffery, 2019; Messer-Misak, de Bruin and Hanke, 2020). Alternatively as has been previously mentioned an adaptive study design pattern could be followed for the overall study design, which would mirror well with the pure Agile development methodology (Pallmann *et al.*, 2018). Overall, given that research into which software development methodologies for medical research is ongoing, we recommend future work consider multiple development methodologies and work closely with the clinical leads to establish contingencies for any changes in requirements.

Fourth, proper training needs to be given to sites for all aspects of a data collection routine and a simple way for them to repeat that training should be developed. Whilst this was performed to a point for PACE in the form of a complete manual this did not appear to be sufficient. One possible solution may be to develop a series of video examples showing how to perform specific functions that sites can access at any time. More generally, future work that develops similar platform for use in multi-site studies may wish to consult closely with senior leaders at both the site and project level and develop a comprehensive training plan using said leader's expert knowledge of their colleagues. Similar approaches are suggested in the literature, ranging from suggestions of consultation with senior leaders to developing specific roles based around technical training and data quality assurance for technical tools (Steinhubl *et al.*, 2019; Rosa *et al.*, 2021)

Fifth, automatic monitoring should be developed as part of standard practice when building similar systems. It would have been trivial to automatically detect missing or incomplete data months in advance of data collection being completed. In particular site 2 clearly had an issue properly using the sensors however this was not picked up until collection had ended. Unfortunately, such software was not implemented as, again naively, it was assumed that the protocol was straightforward enough that few errors would occur. If nothing else, future projects which make use of sensor data should ensure that automatic monitoring is implemented. Automatic monitoring of data is well established in electrical and computer engineering, and is becoming ever more important given the advent of concepts like 'Big Data', 'The Internet of Things' and 'The Ocean of Things' (Bakker *et al.*, 2019; Zhang, Jeong and Lee, 2021). Numerous algorithmic approaches to automatic monitoring have been developed, however one review found that Principal Component Analysis and Artificial Neural Networks are both highly effective for automating such monitoring (Bakker *et al.*, 2019). Alternatively, simpler approaches could preferable in cases such as this study where the amount of data collected is small relative to the levels of data which typically fall under the term 'big data'. For example, in the context of this study data sent via FastFile could be placed manually into a pre-defined folder by the study manager that is monitored by a python script. Once new data is added to the folder, the script can then select the appropriate C3t instance from the database and attempt to locate the timestamps in the data. If the timestamps are found, the data can be plotted and emailed to researchers to confirm its validity. If timestamps are not found the study manager can be notified via email and the site subsequentially contacted. This concept could easily be applied to other similar studies which collect simple sensor data from multiple sites that is easily visualised.

Finally, the ability for clinicians to provide feedback via the app should have been included and focus groups should have been periodically run throughout the app's development. Whilst this would not have directly solved any of the collection issues encountered, it would have provided additional data on which aspects of the software were found difficult to use and led to a more considered user-friendly design. A key theme throughout much of the literature on digital technology for clinical trials points out that technical solutions should be treated first and foremost as collaborations (Steinhubl *et al.*, 2019; Rosa *et al.*, 2021). Thus, it is critical that feedback be facilitated, including after the initial version of any developed software (and hardware) platforms are rolled out to clinicians/end users. Feedback systems in particular are commonplace in professional products, and indeed the concept of Continuous Feedback, a part of Agile development, is based around the idea of evolving software over time in response to user feedback (Ali Babar, 2014). We would suggest that, where possible, bespoke data collection systems developed for research be treated by the developers as closer to a traditional software product, with the various development cycles that come with such products, rather than as 'fire and forget' research tools.

3.6 Limitations

There one particular major limitation of this chapter and the recommendations given - the lack of direct, discussed feedback from the clinicians that used the RDCP (i.e., focus groups). Although clear patterns and problems can be seen in the evaluated data the true feedback should come from discussion with the end users, the clinicians. Unfortunately, such discussions were not included in the time frame of this project. Future work should integrate such evaluation processes into the study design as in order to better understand how future systems can be developed and what pain points there are for end users.

A more minor limitation of the overall app design is the lack of a way for researchers to store the participant's native tongue. The inclusion of this field, which would sit as an enum in the participant object, could potentially be used to spot instances where the native tongue of the respondent was not included in the C3t language options. Given the ease with which such a field could be added, a future update to the app and associated database should include such a field.

3.7 Conclusion

The collection of data is central to all forms of research. As amount of data required for modern analytical techniques increases and the form that data takes changes into more advanced types new collection methods are also required. Whilst remote electronic data collections systems like the RDCP have the potential to drastically improve the way data is collected, handled, and ultimately analysed these systems need to be robustly designed, tested, and monitored for efficacy.

The development of such methods, however, are likely to experience some growing pains, as has been the case with the RDCP.

Unfortunately, there is unlikely to be any simple solution to these growing pains other than trial and error within research groups and organisations. Software development is highly complex and if it is to play a role in medical research needs to be treated as a significant part of the research project. In industry settings software products are routinely designed, tested, and refined over the course of several years by large teams of highly skilled professionals. Research however does not allow for such extended timespans and collection systems must often be ready a matter of months after a project is funded. The need for such rapid development necessitates either the employment of software consultancies which due to their nature can be associated with extremely high costs, or the embedding of software engineers into research organisations. Fortunately, multiple funding bodies are now offering funding for professional research engineers, which will likely go some way towards fixing this problem.

Overall, it is my opinion that the RDCP should be considered a success. The system works as intended and facilitated the collection of significantly more data than would otherwise have been available. Notably the primary data quality issue discussed during this chapter (the poor sensor data quality) can be attributed to the sensor design rather than the RDCP. Whilst the RDCPs performance was not perfect, future projects may be able to learn from the mistakes and assumptions made here in the development of their own systems.

Chapter 4: Exploring the instrumented C3t's relationship with chorea and general motor dysfunction in Huntington's Disease

4.1 Chapter Overview

This chapter focuses on exploring the relationship of sensor features generated from the instrumented C3t to whole-body chorea, upper-body chorea and general motor dysfunction in HD (via the UHDRS-TMS). The relationship of these features with UHDRS-TMS was also included in the analysis to evaluate whether the generated features show a stronger relationship to it than the base C3t time scores, despite their being developed specifically for chorea. The analysis of this chapter was made possible due to the data collected by the RDCP developed as detailed in chapter 3, which provided over 65% of the analysed dataset.

Six objectives are addressed in this chapter of varying clinical and engineering importance. The rationale for each is introduced and discussed in the following section, however for simplicity these objectives are listed below.

Objective 1: Assess the relationship between signal features generated from accelerometers worn during the C3t with clinical measures of chorea.

Objective 2: Focuses on using generating features for objective 1 that are simple to translate into clinical practice.

Objective 3: Assess whether features generated from jerk signals are better for estimating chorea than identical features generated from acceleration.

Objective 4: Assess whether the inclusion/exclusion of the x-axis from generated features had an impact on the feature's relationship with chorea.

Objective 5: Assess whether the features generated from both studied transfer tasks combined (i.e., BTT & CTT) are superior to features generated from a single transfer task (i.e., BTT or CTT).

Objective 6: Assess what impact filter frequency had on the feature's relationship with chorea.

4.2 Introduction

There is an ongoing need for the development of sensitive assessments suitable for monitoring symptoms and their progression in HD (Reilmann *et al.*, 2011a; McColgan and Tabrizi, 2018; Mestre, Busse, *et al.*, 2018). This is both to aid clinical trials, where sensitive measures of progression can show the effect of therapeutics on progression, as well as assist in the clinical management of HD. Whilst composite measures like the CUHDRS have shown increased sensitivity to progression by combining multiple symptom measurements (Schobel *et al.*, 2017), they are still limited by the insensitivity of their component assessments, as discussed in section 1.3.3.8.

This lack of sensitivity is particularly the case with respect to motor symptoms. The UHDRS motor assessment, a series of 31 ordinally rated motor assessments split across 7 assessment areas is routinely used to assess motor symptoms in HD (Kieburtz *et al.*, 1996). The scores from each of these assessments are then summed to produce the UHDRS-TMS, typically considered to be the gold standard measure of general HD motor function (Youssov *et al.*, 2013). As the UHDRS-TMS is the only motor symptom measurement input into the CUHDRS any progression in said symptoms which is not detected by the UHDRS-TMS will not be included in the CUHDRS.

Unfortunately, the sensitivity of the UHDRS-TMS to progression and as an assessment in general is considered to be limited. Whilst multiple longitudinal studies have confirmed that the UHDRS-TMS reliably changes in early manifest HD the same change has not been noticed in pre-manifest or prodromal HD (Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012). However, although progression has been seen in early-HD, the sensitivity of the UHDRS-TMS to that progression is still thought to be limited (Reilmann *et al.*, 2011a; McColgan and Tabrizi, 2018).

This lack of sensitivity is also present for the assessment of specific HD motor symptoms (e.g., chorea, bradykinesia). A common way of assessing specific motor symptoms in HD is to sum the relevant items from the UHDRS motor assessment to produce a single symptom score (Reilmann *et al.*, 2011a). The number of assessments per symptom however varies, ranging from at most 7 assessments (e.g., chorea) to only a single assessment (e.g., bradykinesia) (Kieburtz *et al.*, 1996; Reilmann *et al.*, 2011a). Given that the UHDRS-TMS lacks sensitivity and is made up of all 31 UHDRS motor assessments it seems reasonable to suggest that even coarser assessments will be even less sensitive.

The lack of sensitive motor symptom assessments is not unique to HD. In PD, the gold-standard measure of motor symptoms, the MDS-UPDRS III, uses a similar series of 5-point ordinal scales to assess motor symptom severity (Clarke, 2007). As with the UHDRS-TMS, the MDS-UPDRS III is thought to be coarse with limited sensitivity to change over time (Clarke, 2007). This has led to the widespread development in PD of instrumented clinical assessments which make use of modern sensor technology (e.g., accelerometers, IMUs) (Rovini, Maremmani and Cavallo, 2017). Such assessments have been shown in PD to be sensitive to a number of specific movement disorders such as tremor and gait abnormalities (Rovini, Maremmani and Cavallo, 2017).

The underlying rationale of instrumented assessments is that they are capable of assessing and quantifying motor symptoms directly and objectively. The MDS-UPDRS III and UHDRS motor assessments are based around clinical observation, with expert clinicians having participants perform a range of movements and assessing them for the presence of movement disorders. Whilst expert clinical opinion should not be discounted, there is a limit to the amount of information that can be garnered from them. Measuring properties of movement such as dominant frequencies, entropy, and mean/min/max acceleration/jerk requires sensors.

The counter argument to this is that whilst certain properties of movement cannot be assessed without sensors, the result of them (i.e., the movement disorder itself), clearly can be. The question can be posed that if the cumulative result of these unmeasurable properties, i.e., the movement disorder itself, can be observed then does it matter that its component parts cannot? The problem with this line of thought comes back to the original issue of a lack of sensitivity.

Suppose one attempts to assess the speed of an object, say a car travelling down a road, by rating it as either ‘fast’ or ‘slow’. Inter-rater agreement will likely be very high assuming a common frame of reference (i.e., how fast cars typically travel along a road). As the number of categories expand however it easy to see that inter-rater agreement will likely reduce. At some point, inter-rater agreement will likely become very low, and it will become simpler to directly measure the speed of a car in place of attempting to assess it visually. This is essentially the same rationale for assessing movement disorders using instrumented assessments. Whilst clinical observation is clearly sufficient for assessing motor symptoms there is a limit to how sensitive such scales can reasonably be expected to be. In the context of HD motor assessment, this suggests that if more sensitive assessments are desirable (which they are) then there are two options. First, the scales contained in the UHDRS motor assessment could be expanded, but this will come at the cost of inter-rater reliability past a certain (unknown) point. Second, instrumented assessments could be developed to capture properties of movement directly.

Instrumented assessments have been widely developed for PD however they have not received the same level of attention in HD. Of the instrumented assessments that have been developed for HD they typically focus on the assessment of gait, postural stability, and the UHDRS-TMS (Dalton *et al.*, 2013; Mannini *et al.*, 2015, 2016; Kegelmeyer *et al.*, 2017; Acosta-Escalante *et al.*, 2018; Jensen *et al.*, 2018; Purcell *et al.*, 2019; Gaßner *et al.*, 2020). The exception to this being digitomotography and choreomotography, both part of Q-Motor series of assessments (Reilmann and Schubert, 2017b). As such, there is currently an unmet need for research into instrumented assessments of HD motor symptoms.

The study presented in this chapter seeks to advance the literature by attempting to estimate the severity of chorea seen in gene-positive HD participants using an instrumented version of the C3t. A number of features are extracted from the collected data and their relationship with chorea analysed. The rationale for choosing chorea as opposed to general motor function or a different specific motor symptom is two-fold.

First, there is an initial question when developing any instrumented assessment of motor function as to what aspect of motor function it should seek to assess. In HD there are essentially two options, focus on assessing general motor function using the UHDRS-TMS, or attempt to assess specific motor symptoms using the UHDRS-TMS sub-scores. The difficulty with assessing general motor function using an instrumented assessment is that the underlying motor symptoms which make up the general motor dysfunction are dynamic.

Broadly and typically, HD motor symptom progression can be split into an initial hyperkinetic phase (unintended movement, typically characterised by prominent chorea) which is eventually subsumed by a later hypokinetic phase (poverty of movement, typically characterised by bradykinesia and eventually rigidity) (McColgan and Tabrizi, 2018). From the perspective of measuring general motor dysfunction this presents a problem as two individuals can have exactly the same UHDRS-TMS score but completely different motor symptoms. Thus, an instrumented assessment purporting to be sensitive to general motor dysfunction in HD, which must by necessity show it is related to the UHDRS-TMS, would have to be sensitive to all of the underlying motor symptoms at once. This is further complicated by the fact that the UHDRS-TMS does not weight all symptoms equally (Reilmann *et al.*, 2011a). Chorea for example has seven times the weighting in terms of its contribution to the UHDRS-TMS score than bradykinesia does. As such, there is an inbuilt bias towards instrumented assessments whose measures/features are sensitive to more heavily weighted symptoms (e.g., chorea) when attempting to estimate the UHDRS-TMS. As such, it is simpler to attempt to estimate individual symptoms first and then, if a more general assessment of motor function is necessary, the output measures of multiple such assessments can always be combined later on. It is worth noting that this is in effect exactly what the UHDRS-TMS does – it combines individual assessments of motor function into a single measure of general motor dysfunction.

The second rationale for choosing chorea as the first symptom to assess is due to the role it plays in HD. Historically; chorea has always had a special place in HD. In the original paper 'On Chorea' Dr George Huntington described a type of 'hereditary chorea' which has relatively recently been renamed to Huntington's Disease (Huntington, 1967). Whilst we now recognise that HD is a complex, multi-faceted disease with a wide range of symptoms, chorea is nonetheless a prominent, common symptom (McColgan and Tabrizi, 2018). This is particularly the case in early stages of HD during which substantial, progressive hyperkinetic movements (McColgan and Tabrizi, 2018). Chorea is a particularly common feature during this phase, with a recent review finding of 5609 individuals with clinically manifest Huntington's, the lifetime prevalence of chorea was approximately 96% (McAllister *et al.*, 2021). Importantly many therapeutics under development seek to target the earlier stages of HD during which chorea is likely to be prominent. As such, it makes sense to focus first on developing an instrumented assessment that is sensitive to chorea.

It should be noted that there is currently an instrumented assessment designed to assess chorea in HD included in the Q-Motor series of assessments – choreomotography (Reilmann *et al.*, 2011a). However, whilst choreomotography represented a significant step towards the objective assessment of chorea its performance could likely be improved. The reported correlation with whole-body chorea was at best moderate ($r=0.458$; $p<0.05$), the relationship was not visually shown (Reilmann *et al.*, 2011a), and whether choreomotography truly assesses chorea has been publicly questioned (Casula *et al.*, 2018). On this basis, continued investigation into the development of an instrument assessment of chorea was justified.

The primary objective of this study then was to assess the instrumented C3t's (henceforth distinguished from the non-instrumented version of the C3t as the *instrumented C3t*) ability to estimate whole-body and upper-body chorea as measured by the UHDRS motor assessment. Specifically, sensor data collected from two accelerometers worn on the wrists whilst the C3t was performed is utilised. To effectively determine the efficacy of the instrumented C3t in this regard, it is useful to have a baseline against which to judge success. The natural task to compare the instrumented C3t's perform against is the base C3t time scores. However, in chapter 2 it was shown that the non-instrumented C3t does not have an observable monotonic relationship with whole-body or upper-body chorea. Whilst this means that any such relationship using the instrumented C3t would be an improvement, a more stringent success criteria is required. The other natural comparator is the choreomotography assessment the performance of which will be used here to judge the instrumented C3t's efficacy. Additionally, the non-instrumented C3t is known to be highly correlated with (and can be used to accurately estimate) the UHDRS-TMS (Clinch *et al.*, 2018; Woodgate *et al.*, 2021). As such, the relationship of the instrumented C3t with the UHDRS-TMS is included in the analysis presented here to determine if the instrumented C3t shows improved performance relative to the non-instrumented C3t.

There are a number of additional objectives regarding the development of the instrumented C3t as an assessment of chorea that this chapter also addresses.

In keeping with the discussion in chapter 1, an underpinning aim of this thesis is to focus on the development of features from instrumented assessments that are simple to translate into clinical practice. As such, the second objective of this study is to generate features from the instrumented C3t related to chorea to which simple clinical meaning can be attributed.

The third objective of this study is to explore whether jerk, the first derivative of acceleration, is a better estimator of chorea than acceleration. Accelerometers were used in this project as they were felt to be suitable sensors for assessing movement disorders as they can record the movements participants make as they take the C3t. The acceleration of the sensors during the instrumented C3t may be linked to chorea with random choreatic movements presumably resulting in sudden increases in acceleration. These sudden increases in acceleration however may be better identified by looking at the rate at which they occur, and jerk has been shown before to be related to motor symptoms in PD (Eager, Pendrill and Reistad, 2016; Rovini, Maremmani and Cavallo, 2017). Thus, jerk signals are analysed in addition to the acceleration signals to see if they offer any performance increase. Identical features are extracted from both acceleration and jerk signals and their relationship to chorea (and for completeness the UHDRS-TMS) is compared.

The accelerometers used in this study are tri-axial, meaning the record acceleration along three distinct axes (x, y, z). The majority of the intentional movement during the C3t occurs along the x-axis (see section 1.3.2.2). Thus, the unintentional movements caused by chorea may be better isolated by only utilising data from the y- and z-axes where they will be presumably less obscured by intentional movement. Thus, the fourth objective of this study is to assess whether features generated from the y- and z-axes are superior to those generated from the x-axes in terms of their relationship with chorea and the UHDRS-TMS.

The fifth objective of this study is to determine whether information from both the BTT and CTT are required to assess chorea. Ideally, only one task would be needed for the same reasons discussed in chapter 2 – the simpler the instrumented C3t is to conduct the easier it can be widely applied in large-scale clinical studies and potentially in the home. To assess this, features are first generated using both tasks, and then the best performing ones will be re-generated using information from only a single task, allowing their performance to be compared.

Finally, the sixth objective of this study is to determine whether there is an optimal filter frequency range for assessing chorea. In PD, the frequency of tremor has been widely reported (Deuschl, Bain and Brin, 1998) and as such features generated within this frequency range should be more sensitive to tremor as a result. The optimal frequency range, if there is one, for assessing chorea however is unknown. As such, a heuristic search for an optimal frequency band for the assessment of chorea is conducted in this study.

4.3 Methods

4.3.1 Methods Overview

This section is split into two parts - section 4.3.2 details how the data was collected, and section 4.3.3 covers how data was processed and analysed. Each section is then further divided into subsections as required, with background information and technical details given throughout. Importantly, as this is an interdisciplinary project background information about methods used is provided as what may be considered to be 'common knowledge' in one field may not in others.

4.3.2 Data Collection & Experimental Setup

4.3.2.1 Participants

As in chapter 2, data used in this study were drawn from multiple other studies – PACE-HD, TRIDENT and Developing Clinical Applications for a Novel Multi-Task Functional Assessment: The Clinch Token Transfer Test (referred to here as C3t PhD). For information on each of the studies see section 2.3.1.1.

Participants were subdivided into different disease stages in the same manner as in section 2.3.1.1 using the TFC and Diagnostic Confidence Level (DCL), as shown in Table 38.

Table 38: HD disease stage DCL and TFC requirements

Disease Stage	Requirement	Broad description
Pre-manifest	DCL < 1	Yet to manifest over motor symptoms
Prodromal	DCL = 2-3	
TFC Stage 1	DCL = 4; TFC = 13-11	Manifest HD, symptomatic
TFC Stage 2	DCL = 4; TFC = 10-7	
TFC Stage 3	DCL = 4; TFC = 4-6	

Notably, the analysed dataset did not include control participants. The rationale for this was two-part. First, the conducted analysis requires UHDRS motor assessment scores which were not available for any control participants who had instrumented C3t data. Secondly, the overall goal of this study was to assess whether features generated from the instrumented C3t could be used to estimate the severity of chorea in HD, not whether they could distinguish between control and HD populations. Future work using a control population should however be conducted, as is discussed in section 4.6.7.2.

It should be noted that the inclusion of participants at different clinical stages of the disease was necessary both to secure a pool of participants with various levels of chorea severity, and as a practical step to not hinder data collection and therefore potentially limit sample size. As the different manifest stages are not dependent on motor symptom progression it was felt that doing so would not bias the results of this study. Additionally, this approach has been followed in similar and related work (Reilmann *et al.*, 2011b; Bennasar *et al.*, 2018; Clinch *et al.*, 2018). However, whether there is any relationship between the generated features and the different clinical stages should be explored in future work, as is also discussed in section 4.6.7.

Finally, it should be noted as well that previous work has already shown that sensor features generated from sensors worn during the C3t can be used to distinguish between control and HD participants (Bennasar *et al.*, 2018).

4.3.2.2 C3t & Accelerometer Data

All participants performed either the C3t or its earlier version, the MBT, during a single visit using the C3t android app to administer the test. As again in chapter 2, both tasks will be referred to as the C3t from here on.

Whilst taking the C3t participants wore one tri-axis GeneActiv accelerometer (Activinsights, UK) on each wrist (dominant hand & non-dominant hand) set to record at 100Hz. As mentioned in section 1.5.2.3.3.3, GeneActiv accelerometers were chosen for this study as a previous study run with the C3t had collected GeneActiv data (Bennasar *et al.*, 2018), and data from this prior study could be pooled with data collected during this study, and so increase the available sample size. As shown in Figure 34, on the dominant hand the sensors were worn with the connector pins facing down and on the non-dominant with the connector pins facing upwards. This was to allow for which sensor belonged to which to be distinguished in the dataset even if the handedness of the participant was unknown.

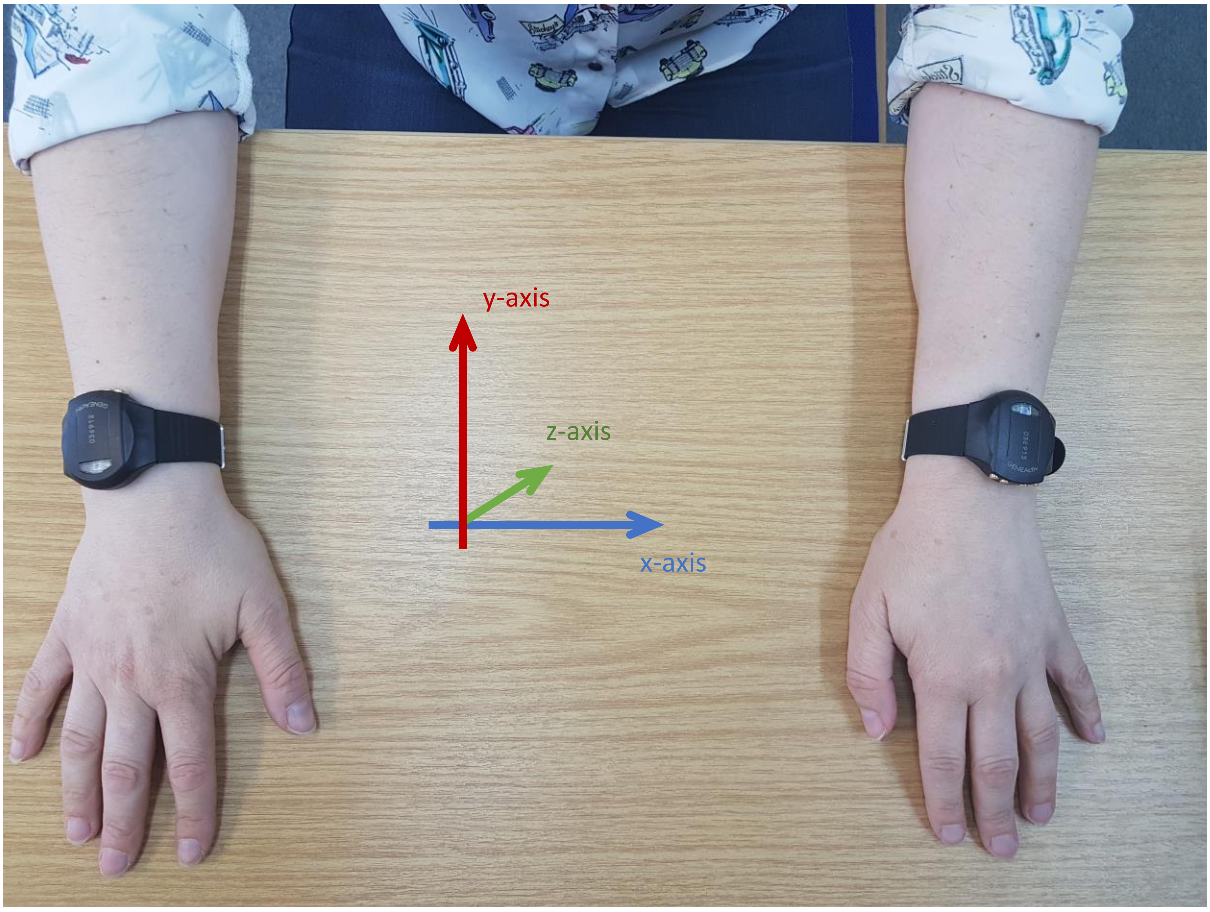


Figure 34: GeneActiv accelerometers worn during the C3t. Note the placement of the pins on the accelerometers are reversed, allowing which sensor was on which hand to be distinguished if the handedness of the participant was not known.

The axes of the watches correspond to movement along the C3t test board as follows:

- The x-axis records motion across the board (width)
- The y-axis records motion up and down the board (depth)
- The z-axis records vertical motion towards and away from the board (height)

As in chapter 2, to ensure compatibility between the C3t and MBT datasets, only data from the Baseline Transfer Task (BTT) and Complex Transfer Task (CTT) are used in this study. As it was found in chapter 2 that neither the study site nor the test version contributed to test performance the datasets were considered valid to merge.

Each participant had twelve acceleration signals recorded from:

- 3 acceleration signals per accelerometer (x-, y-, z-axes)
- 2 accelerometers per task
- 2 tasks

All data were collected using the RDCP described in chapter 3 with the exception of 9 participants from the C3t study which were collected before the RDCP was developed. As was the case in chapter 2, significantly more data was available for the BTT and CTT tasks. As such only data from these tasks is analysed here.

4.3.2.3 Clinical Scores

Each recruited participant also completed the full UHDRS assessment battery within 6 months of taking the C3t. Three measures of motor symptoms were extracted – the UHDRS-TMS, whole-body chorea and upper-body chorea. The measures of whole-body and upper-body chorea were calculated in the same manner described in section 2.3.1.2.2 – i.e., by summing the individual chorea items as shown in Table 39.

Table 39: Chorea assessment areas from the UHDRS motor assessment and their inclusion/exclusion from general measures of whole- and upper-body chorea

Chorea Assessment Area	***Whole-body chorea***	***Upper-body chorea***
Head	*x*	*x*
Face	*x*	*x*
Trunk	*x*	*x*
BOL	*x*	
Left-upper limb	*x*	*x*
Right-upper limb	*x*	*x*
Left-lower limb	*x*	
Right-lower limb	*x*	

4.3.3 Data Analysis

4.3.3.1 Overview

The analysis conducted in this study consisted of three steps – data pre-processing, feature generation, and statistical analysis. Pre-processing refers to extracting the accelerometer data and getting it to a point that it can be analysed. Feature generation refers to taking the pre-processed accelerometer data and calculating discrete quantities (a.k.a. sensor features) from it that may be related to chorea. Finally, statistical analysis was performed to determine the relationship between the sensor features and chorea & the UHDRS-TMS.

4.3.3.2 Data Pre-processing

4.3.3.2.1 Pre-processing: Overview

The goal of data pre-processing stage was to take the raw acceleration signals recorded during the C3t and turn them into signals suitable for extracting & engineering features from. As a reminder, each participant ultimately had 12 acceleration signals recorded for them (3 signals per sensor, 2 sensors per task, 2 tasks). At the end of the pre-processing stager each signal had been pre-processed in the following manner. The steps taken were as follows.

1. The raw signal was extracted from the sensor and segmented into its BTT and CTT components

2. The extracted signals were filtered using four bandpass filters with a low pass cut off of 0.3Hz (chosen because nearby electrical cables can cause interference at 0.3Hz and below (Martinez-Manzanera *et al.*, 2016)) at and a variable high pass cut off of 20Hz, 13Hz, 7.5Hz and 3Hz, resulting in four copies of the signal being created

3. Each of the acceleration signals was then converted into a jerk signal, using 10ms (the smallest difference possible as the sensors records at 10ms intervals i.e., 100Hz) as the time difference between signal samples

4. Each acceleration signal was normalised with respect to time, converting each signal in 1% chunks with a 50% overlap between chunks, making each signal consist of 200 samples

Each step is detailed in the following subsections. Each of these steps is visually displayed in Figure 35 and an example of the outputs from each pre-processing stage is shown in Figure 36.

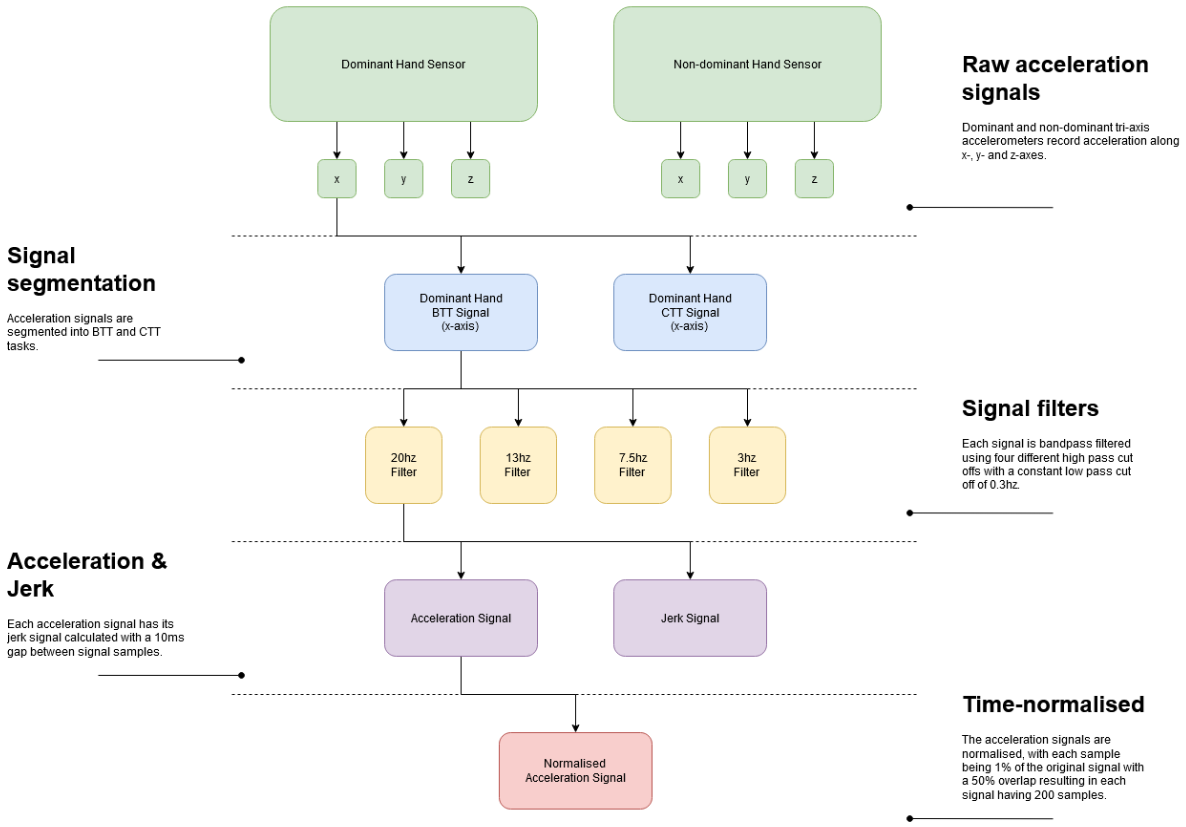


Figure 35: Pre-processing accelerometer signals step-by-step for a single run-through of the x-axis. All processing steps are applied to all axis of both sensors individually. The arrows display all possible options at a given step, the trace/path of a single walk-through being shown all the way to the bottom (i.e., x-axis, dominant hand, BTT task 20Hz filter acceleration signal, normalised).

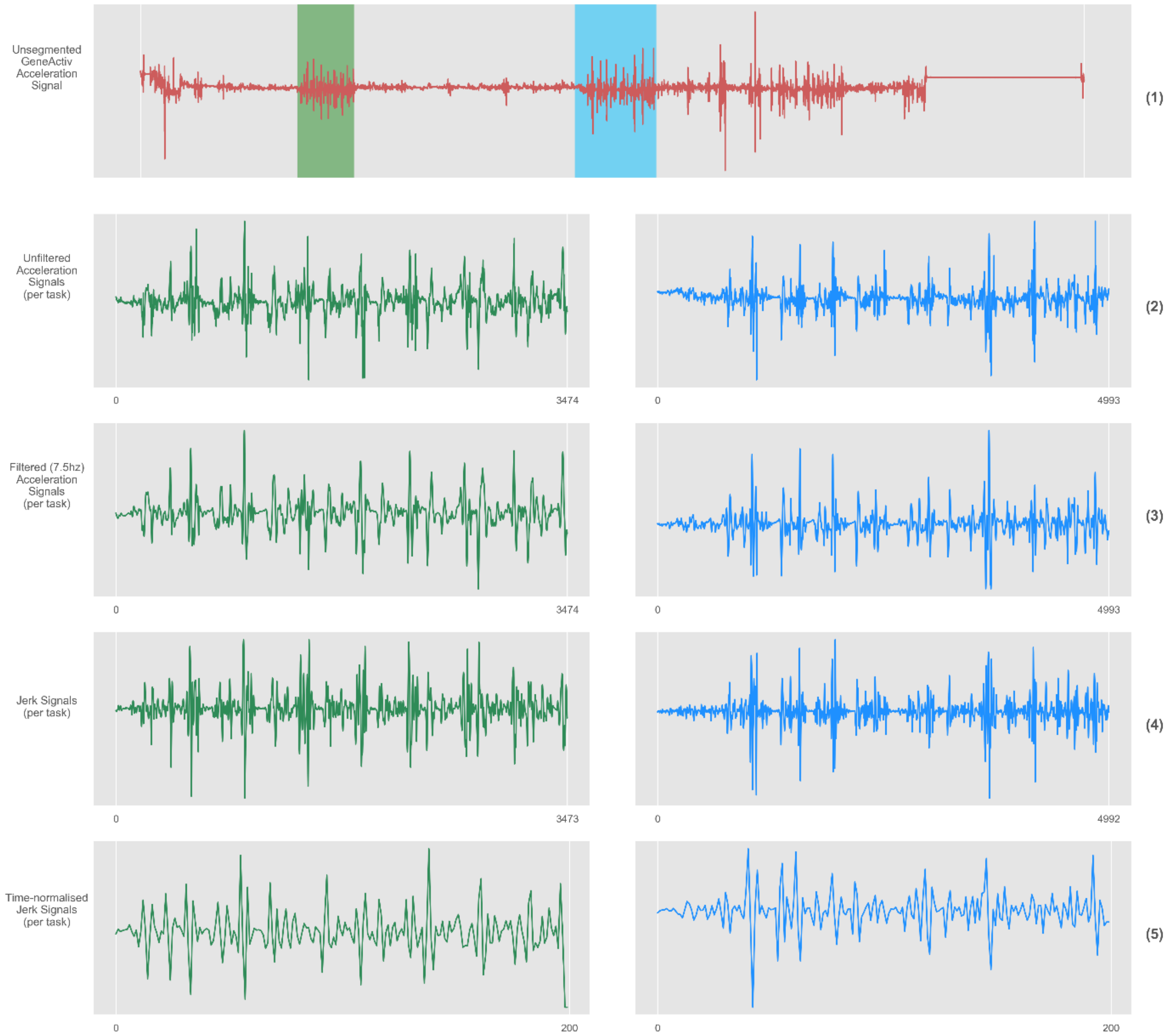


Figure 36: Example of each step in the pre-processing stage for an acceleration signal. Green signals are from the BTT, blue signals are from the CTT. (1) shows the raw acceleration signal recorded during the whole C3t with the BTT and CTT highlighted in green and blue, respectively. (2) Shows the segmented acceleration signals for the BTT and CTT. (3) Shows the BTT and CTT acceleration signals after filtering with a high-pass filter of 6.5Hz and a low-pass filter of 0.3Hz. (4) shows the filtered BTT and CTT acceleration signals converted to jerk. (5) shows the BTT and CTT jerk signals time normalised. The x-axis on each graph shows the first and last sample number.

4.3.3.2.2 Pre-processing step 1: Segment the acceleration signal

As the GeneActiv accelerometers recorded the entire C3t assessment often including substantial amounts of time either side of the assessment, it was necessary to first segment the acceleration data into individual C3t tasks. Signals were segmented by matching the timestamps recorded on the C3t android app with those recorded by the accelerometers. To account for slight differences between when the participant was instructed to start and/or finished the assessment and the time the clinician recorded them doing so, one second was added either side of the reported start/finish times.

An example of the segmentation process is shown in Figure 36 where (1) shows the unsegmented signal and (2) shows the segmented BTT (left) and CTT (right) tasks, respectively.

Two algorithms were used during the segmentation process, one for converting between Unix and GeneActiv timestamp formats and one for determining whether a GeneActiv file contained the correct timestamps. These two algorithms are described below.

4.3.3.2.2.1 Timestamp conversion algorithm

Timestamps were recorded differently on the C3t app and GeneActiv accelerometers. The C3t app in standard Unix millisecond format, i.e., the number of milliseconds since January 1st, 1970, UTC-0. The accelerometers record timestamps in the following format 'year-month-day hour:minute:second:millisecond' with the time-zone depending on the location of the device.

For example, at time of writing the date and time is 2020-06-08 12:45:30:000 which in milliseconds would be 1591616730000. Conversion of the Unix timestamp to the GeneActiv format was accomplished using the following algorithm.

1. Extract time-zone information from GeneActiv file
2. Increment the Unix timestamp by 3,600,000 multiplied by the time-zone offset (i.e., UTC+2 would be 3,600,000 x 2)
3. Convert the number milliseconds into GeneActiv format using
4. If the Unix timestamp was the start time of a task, round the number of milliseconds *down* to the nearest 10 milliseconds
5. Else if the Unix timestamp was the finish time of a task, round the number of milliseconds *up* to the nearest 10 milliseconds

4.3.3.2.2.2 Segmentation Algorithm

Due to the size of some of the GeneActiv files directly reading the whole file would be at best inefficient and at worst impractical. However, as the recording speed to the accelerometers is known (100Hz), the offset between the first recorded sensor reading and the time a task started could be calculated. This allows the line number of the start and finish time of a task to be calculated and then those specific lines along with the ones in between read, skipping the rest.

The algorithm used to achieve this and so determine whether a given file contained the correct timestamps and subsequently segment the signal into tasks was as follows.

1. Collect the start and finish Unix timestamps for a given task
2. Read the time-zone modifier from the GeneActiv file
3. Adjust the Unix timestamps based on the time-zone modifier

4. Convert the Unix timestamps into GeneActiv format as previously described
5. Read the first GeneActiv sensor reading timestamp
6. Calculate the difference between the first timestamp and the task start timestamp
7. Calculate the difference between the task start timestamp and task finish timestamp
8. Jump to the calculated start and finish index in the file, segment and return the task signal if both indices were found, else notify of failure

4.3.3.2.3 Pre-processing step 2: Filter the acceleration signals

A common pre-processing step in signal analysis is to filter the signals before analysing them with the goal of removing noise from the signal. An example that shows the effect of bandpass filtering the BTT and CTT acceleration signals (high-pass=7.5Hz; low-pass=0.3Hz) is shown in Figure 36, where (2) shows the unfiltered signals and (3) shows the filtered signals.

Filtering refers to the process of converting a time-domain signal into the frequency domain, removing certain frequencies, and then converting it back into a time-domain signal. Filtering approaches are typically referred to as one of ‘high-pass’ (frequencies *above* a threshold are removed), ‘low-pass’ (frequencies *below* a threshold are removed) or ‘bandpass’ (high & low-pass filters applied, creating an allowed frequency band).

Human motion is known to not exceed 20Hz (Bouten *et al.*, 1997) and at 0.3Hz & below nearby electrical cables may interfere with sensor recordings (Martinez-Manzanera *et al.*, 2016). As such, a basic approach to filtering acceleration signals generated from human motion is to bandpass filter the signals with a high-pass cut off of 20Hz and a low-pass cut off of 0.3Hz.

Using such an approach however whilst valid is not necessarily optimal. Tremor for example is known to operate within 4-6Hz (Deuschl, Bain and Brin, 1998) and thus a bandpass filter of 20Hz to 0.3Hz would likely result in noisy data for the purpose of identifying and assessing tremor.

Unlike tremor, the optimal frequency band for identifying chorea is unknown. As such observing the effect (in terms of correlation strength & predictive performance) different filter frequency bands have on the generated features could provide insight into what an effective filter frequency for chorea may be.

To determine the effect of different frequency bands the raw acceleration signals were filtered using four different high-pass filters. Thus, the raw acceleration signal was turned into four distinct signals each of which had a different frequency range removed. The same features were then calculated for each of these signals and, by looking at the difference in performance between them in terms of correlation strength and prediction quality, the impact of different high-pass filters was determined.

Each acceleration signal was bandpass filtered using a 2nd order Butterworth bandpass filter (in line with previous studies (Martinez-Manzanera *et al.*, 2016)) with a low-pass cut off of 0.3Hz and a variable high-pass cut off. The high pass cut offs used were as follows:

- 20Hz
- 13Hz
- 7.5Hz
- 3Hz

The high pass filters other than 20Hz were selected using a heuristic approach based on the signals spectral edge frequency (see below). After this process was completed, each participant had 48 signals (12 per filter frequency).

4.3.3.2.3.1 Selecting high-pass filter cut offs using spectral edge frequency

As the optimal frequency band for identifying chorea is unknown a heuristic approach was taken to find a series of candidate high-pass frequency cut offs. The general idea was to calculate the spectral edge frequency for each signal (the frequency at which some percentage of the power of a signal is contained) and then to base the selection of high pass cut offs on the identified spectral edge values.

Each spectral edge frequency of a signal was calculated after that signal had been filtered with a high-pass cut off of 20Hz and low-pass cut off of 0.3Hz. This was to remove any high/low frequency noise from the signal which might affect the calculation of the spectral edge frequency.

Specifically, the following process was conducted.

1. The signals were filtered using a 2nd order Butterworth bandpass filter (high-pass 20Hz, low-pass 0.3Hz)
2. The 99th, 95th and 75th spectral edge frequencies were calculated for each filtered signal, across all participants, axes, sensors, and tasks
3. For each spectral edge frequency, the mean value was then calculated and used as a high pass cut off, rounded to the nearest 0.5Hz

4.3.3.2.4 Pre-processing step 3: Convert the acceleration signals into jerk signals

As stated in section 1.5.2.3.4 and section 4.2, it is possible that jerk may be more sensitive to choreatic movements than acceleration. Jerk is the first derivative of acceleration with respect to time calculated using the following equation where $\vec{j}$ is jerk and $\vec{a}$ is acceleration (Eager, Pendrill and Reistad, 2016).

$$\vec{j}(t) = \frac{d\vec{a}(t)}{dt}$$

The rationale for observing the jerk signal is rooted in the clinical description of chorea. Chorea is described as *"involuntary, unpredictable jerk-like muscle contractions that randomly involve different body parts and vary in frequency, intensity, and amplitude"* (Jankovic and Roos, 2014). Thus, a single choreatic movement will result in the sudden acceleration of the affected area, potentially producing a follow-up movement in other areas. In terms of an acceleration signal, this would be expected to result in a sudden increase in acceleration. The rationale for assessing chorea using jerk should now be evident – a sudden increase in acceleration is described in terms of its change relative to time, jerk.

All acceleration signals were converted into jerk signals using the equation listed above. The smallest possible difference between signal samples was used to calculate jerk which, as the recording speed of the accelerometers was set to 100Hz, was 10ms.

4.3.3.2.5 Pre-processing step 4: Normalise the acceleration & jerk signals with respect to time

The final pre-processing step was to normalise each of the signals with respect to time. The rationale for doing so was two-fold.

First, based on the findings of chapter 2, we can surmise that the time each task takes appears to be influenced by multiple factors only one of which is motor symptoms. As the time taken increases, so too will the length of each acceleration signal. The difficulty this poses for attempting to estimate chorea is that it is not just chorea that causes the increase. Thus, any generated feature whose value is influenced by the length of the signal may potentially influenced by multiple factors, with chorea being only one of them. Whilst features generated from time-normalised signals will still be influenced by the number of samples present in the underlying signal, the effect should at least be reduced.

Second, the recording speed of 100Hz will likely generate a good deal of noisy data where little change occurs. The longer the signal the more of this noise will be present. A typical approach to dealing with such noise is to convert the signal into epochs, for example 1 second epochs, as is done by Bennasar *et al.*, (2018a). Normalising the signal with respect to time will however accomplish the same thing - removing the noisy data whilst retaining the large structures thought to be relevant to chorea. Essentially, there is no reason to treat 1 second as a magic number, in this study it was felt that 1% made more sense. A similar approach is routinely used in gait analysis, with recordings of multiple participants performing the same walk being time-normalised such that the signals contain the same number of samples and so their structures can be compared (Whittle, 2007).

Specifically, in this study each signal was converted into 200 samples where each sample was 1% in length of the original signal and overlapped its preceding and successive samples by 50%. Thus, a signal with 1000 samples would be converted to a 200-sample signal, where each sample contained 10 samples, 5 of which are included in the previous sample and 5 of which are included in the following sample.

In Figure 36 an example of the result of this process is shown, with (4) showing the original jerk signal and (5) showing the time-normalised version.

4.3.3.3 Feature Generation

4.3.3.3.1 Feature Generation: Overview

The goal of the feature generation stage is to create features from the acceleration & jerk signals that may be related to chorea in HD. There are three steps to this process.

1. Features thought to be useful for assessing chorea are defined (feature definition)

2. Features are extracted on a per-signal basis (feature extraction)

3. Where appropriate, features from individual signals are combined to produce single composite features (feature engineering)

Each of these steps is covered in turn in the following subsections.

4.3.3.3.2 Feature Generation step 1: Feature Definition

The primary objective of this study was to attempt to assess whole-body and upper-body chorea in HD with a secondary objective that the features extracted must be simple to translate into clinical practice. As such, the first step during feature generation was to consider what properties of the acceleration and jerk signals might be related to chorea and also fit naturally with the clinical description of chorea.

As stated in section 4.3.3.2.4, chorea is described as *"involuntary, unpredictable jerk-like muscle contractions that randomly involve different body parts and vary in frequency, intensity, and amplitude"* (Jankovic and Roos, 2014). As was also stated in section 4.3.3.2.4, sudden, random movements could be reasonably expected to produce sudden increases in acceleration. Such peaks could also be reasonably be assumed to result in spikes in the corresponding jerk signal. In terms of signal properties, a random increase in acceleration would presumably generate a peak in the signals. Parallel to this, the UHDRS motor assessment rates chorea as worsening when it occurs more commonly and with greater degree of severity (Kieburtz *et al.*, 1996). Thus, it was thought that two simple features could be generated from the signals which might be related to chorea – the number of peaks in a signal and the mean width between those peaks. The thought was that as chorea worsens the number of peaks in the acceleration/jerk signal would similarly increase, and the distance between these peaks would decrease. These features are described below in Table 40.

Table 40: Initial feature descriptions

Feature Name	Feature Description
Peak count	The number of peaks in a signal, where a peak is defined as a point in the signal where it is higher/lower than the preceding value and lower/higher than the following value.
Mean peak width	The mean width between all peaks in a signal, where a single width is defined as the number of samples between two peaks where one peak occurs directly before/after the other

It should be noted that the peak and width features defined here are only two of many features which could have been generated. Examples of commonly used features generated from acceleration signals include the root mean square (which has been used to assess gait in PD (Ferrari *et al.*, 2016)), sample entropy (which has been used show differences between essential tremor and PD patient groups (Ruonala *et al.*, 2014)), and the recurrence rate (which has been used to distinguish between control and HD groups (Bennasar *et al.*, 2018)). Similarly additional features of the peaks could have been looked at here (e.g., mean/max/min/x-percentile peak height, mean/max/min/x-percentile peak width, etc).

However, as was discussed in section 1.4.3.1, and is listed as objective 2 of this chapter, a focus of this thesis was to keep the features utilised here to be as few and as simple as possible. As was also discussed in section 1.4.3.1 the rationale behind this was to simplify clinical translation, avoid 'mining' the data, and reduce the impact of The Curse Dimensionality & type I errors (see section 4.3.3.3.4.1). Overall, whilst many potentially useful features could be generated from the collected acceleration data, we deliberately limited our investigation to the features defined above. Exploring additional features of the collected data and their connection to various aspects of HD should however be the focus of future research and is listed as such in section 4.6.

4.3.3.3.3 Feature Generation step 2: Feature Extraction

Once the features to be extracted are decided on the next step is to actually extract them. The *find_peaks* and *peak_widths* functions from the SciPy Python extension were used to calculate the number of peaks and widths in a signal respectively (Virtanen *et al.*, 2020).

As the acceleration and jerk signals can take on both positive and negative numbers, and the direction of the movement was not relevant, both peaks and troughs were considered to be valid peaks. The full width between peaks was used as the width between peaks. The number of peaks and mean width between them was calculated for every individual signal and the results stored in a local SQLite database.

4.3.3.3.4 Feature Generation step 3: Feature Engineering and final feature definition

During the data pre-processing stage multiple different signals were extracted for analysis (see section 4.3.3.2). For every filter used, a given participant had a total of 48 distinct signals (2 tasks, 2 sensors per task, 3 axis per sensor, 2 sets of signals (acceleration & jerk)) from which features could be generated. As the acceleration and jerk signals were to be analysed in isolation, this means that for a given feature 24 signals could be analysed simultaneously. As such, this meant that even with only two features, for each signal type a total of 48 features would be generated.

Generating a high number of features is common in modern data analysis. It does however present a problem for building statistical models and making inferences about the dataset, commonly referred to in machine learning literature as The Curse of Dimensionality.

4.3.3.3.4.1 The Curse of Dimensionality

The Curse of Dimensionality is a term coined by R. Bellman in 1957 and refers various phenomena that occur when working with high-dimensional data that do not occur when working with low-dimensional data (Bellman, 1957). In the context of feature generation, high-dimensional data (i.e., a large number of features) presents two specific problems.

The first problem is that as feature space increases so does the required sample size for prediction models to be meaningful and robust. When the number of features is significantly higher than the number of samples (commonly noted as p>>n), models can overfit to the data and so produce unreliable, overoptimistic results (Vabalas *et al.*, 2019). One 'rule of thumb' recommendation is that for every feature there should be at least 5 samples (Sergios Theodoridis and Koutroumbas, 2009). Note that this recommendation is relevant to multivariate models (i.e., models which combine multiple features together) rather than multiple models with a single variable each.

The second problem is that the large number of features makes it difficult to give meaning to generated models. Importantly in this study there were essentially two options for analysing the data – either every feature could be analysed in isolation, or they could all analysed together, and feature selection employed to reduce the dataset down to the 'best' features. The issue with either approach from a clinical explanation standpoint is that in either case a rationale would have to be given as to why some features were preferred over others. It would be difficult for example to explain why '*the number of peaks in the jerk signal of the z-axis of the dominant-hand sensor during the BTT*' was more important the same feature but from the non-dominant sensor outside of random chance.

There are two common methods of addressing the curse of dimensionality – feature selection and dimensionality reduction.

Feature selection refers to the process of selecting 'good' features from a training set of data and applying them to a testing set of data. However, as Vabalas *et al.*, (2019) points out, feature selection if not performed properly can lead to significant bias in model performance - a 'nested' approach to feature selection is required for it to be performed properly. Unfortunately, following a nested approach to feature selection is problematic with small datasets as the nested training-testing splits can become very small. Additionally, feature selection does not remove the initial issue noted that clinically explaining why a small subset of highly specific features are together sensitive to chorea would likely be difficult.

Dimensionality reduction refers to taking the whole feature space and reducing the number of features by combining them together (Sorzano, Vargas and Montano, 2014). Dimensionality reduction has the advantage here of making use of all the original features but combining them such that only a single variable per feature type is ultimately analysed. Whilst there are numerous techniques for performing dimensionality reduction (Sorzano, Vargas and Montano, 2014), a simple approach is taken here again rooted in the current clinical understanding of chorea.

Choreatic movements are described as *"involuntary, unpredictable jerk-like muscle contractions that randomly involve different body parts and vary in frequency, intensity, and amplitude"* (Jankovic and Roos, 2014). The chorea component of the UHDRS motor assessment rates chorea in increasing severity as *"absent" "slight/intermittent", "mild/common or moderate/intermittent", "moderate/common", and "marked/prolonged"* (Kieburtz *et al.*, 1996).

When attempting to observe a random phenomenon, which chorea effectively is, there are essentially two options – create conditions such that the likelihood of the event occurring increases or increase the sample size of the recorded observation such that the phenomenon is likely to be captured.

Translated to observing chorea in tri-axial accelerometer signals, this means either the chance chorea occurs needs to be increased (which we cannot) or else increase the amount of data we under observation such that when chorea does occur, we will notice it and, ideally, get an idea of how regularly occurs.

As such, attempting to measure chorea by just observing a single axis is likely to be insufficient. A better approach may be to combine features from multiple axis, sensors, and tasks together in order get an idea of how a participants' signals typically look. Combining features in this manner not only makes sense clinically, but also helps to solve the high-dimensional feature space problem.

4.3.3.3.5 Final Feature Definition

Two types of features were generated per *signal* – the number of peaks and the mean width between them. These features were then combined together as follows.

For each feature, the mean value was calculated from both sensors and both tasks. The number of peaks was then also divided by 200, the number of samples in time-normalised signals, converting it into the mean ratio of peaks per signal. As stated in section 4.2, one consideration was whether the purposeful movements during the C3t which occur predominantly along the x-axes would obscure the random purposeless movements created by chorea. As such a set of features was generated by combining the x-, y- and z-axes and a second set with just the y- and z-axes.

Ultimately 32 distinct composite features, 16 per feature type (peak count & mean width), were created and analysed. Figure 37 shows the pathway followed for their construction.

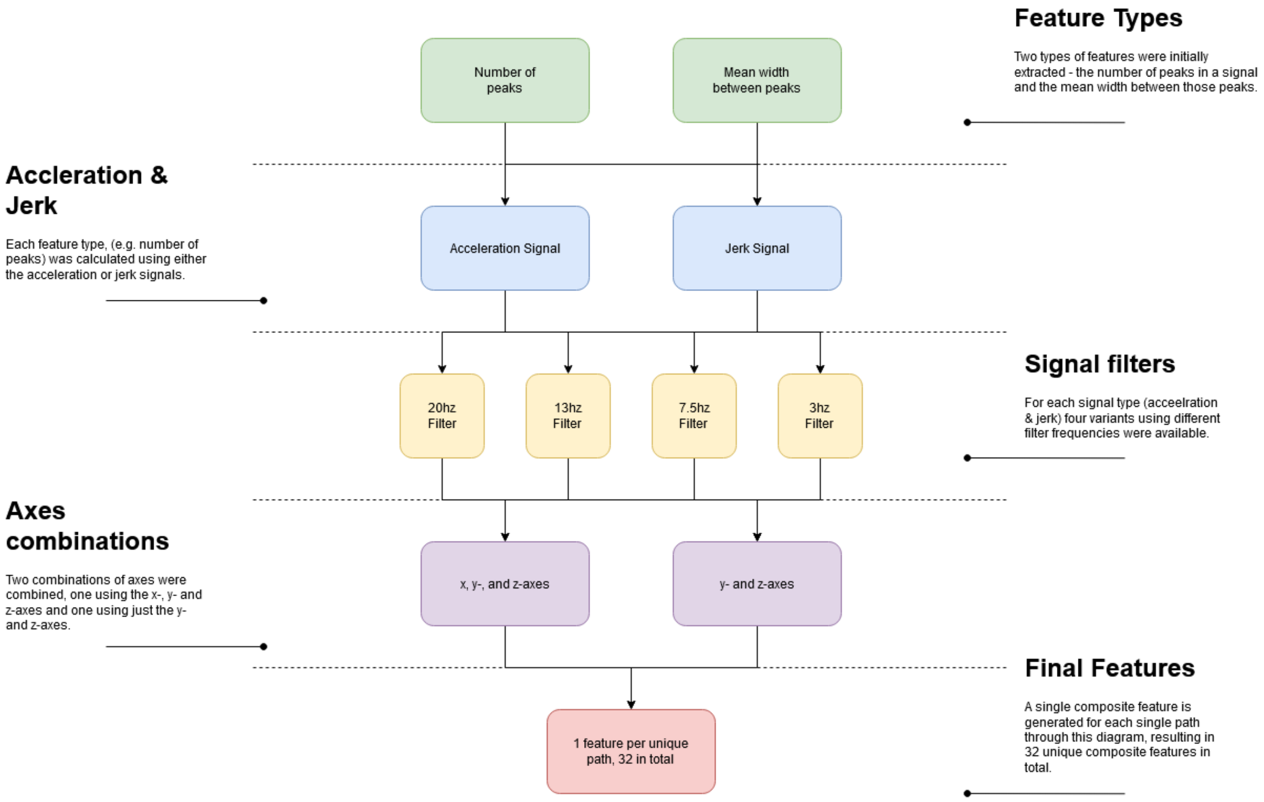


Figure 37: Pathways of feature generation. Two features, number of peaks and mean width between peaks, were calculated per signal. These were then combined by calculating the mean between groups of features calculated across related signals. A single trip through the pathways above from top to bottom fully describes a single unique feature. For example, one such composite feature is the number of peaks, calculated from the acceleration signal, filtered at 20Hz, across the x-, y- and z-axes. This composite feature would consist of the mean number of peaks from 6 signals with the profile of those signals described by the pathway.

4.3.3.4 Statistical Analysis

4.3.3.4.1 Statistical analysis overview and analyses-objective breakdown

The analysis performed to address the objectives outlined in section 4.2 was split into two parts.

To address objective 1, the relationship between the composite features generated from *both* the BTT & CTT and the clinical measures of whole-body chorea, upper-body chorea, and the UHDRS-TMS were assessed. Using the information from objective 1 various best performing features, in terms of their correlation strength with chorea, along with specific variants of them were extracted. Differences between these features were then assessed to in order to address objectives 3, 4, 5, and 6.

The following subsections are split into two parts. Section 4.3.3.4.2 describes the analysis conducted to achieve objective 1. Section 4.3.3.4.3 uses the information from Part 1 to address objectives 3, 4, 5, and 6. Each of the analysis steps are described in the order they were performed. Objective 2 is implicit in the design of the features and so the suggested explanation of the most effective features is discussed in section 4.5.

4.3.3.4.2 Statistical analysis part 1: Assessing feature relationship with chorea and the UHDRS-TMS

4.3.3.4.2.1 Histograms, scatter plots and normality

Each of the 32 features were initially assessed visually. As in chapter 2, visual analysis of histograms was used to assess the distribution of the sensor features and remove any which were found to be invariant. Scatter plots were then used to assess whether retained features showed a monotonic relationship with the clinical measures of whole-body and upper-body chorea.

The normality of each feature retained for analysis was assessed using histogram, Q-Q plots, and three statistical tests (D'Agostino K-Squared, Anderson-Darling, and Shapiro-Wilks). As the majority of scores were found to be non-normal, non-parametric statistical methods were preferred.

4.3.3.4.2.2 Correlation & regression analysis

All combined sensor features found to have a monotonic relationship with whole-body and upper-body chorea were assessed using correlations and regression.

Spearman's R was used to determine the correlation strength between the features and each clinical measure. Significance was initially assumed if $p<0.05$. As in chapter 2, due to the multiple comparisons being made Holm-Bonferroni corrections were applied post-hoc to reduce the chance of type I errors.

Ordinal linear regression was then used to assess the ability of each feature to predict the chorea scores. Again, as in chapter 2, K-fold cross validation ($k=5$, repeats=10) was used to reduce the chance of overfitting the dataset and MAE & Normalised MAE used to assess model quality.

Note that as discussed in section 4.3.3.3.4.1, multivariate analysis would not have been suitable for the available data due to the high number of features relative to the number of samples. As such ordinal models were produced *per feature* and the Holm-Bonferroni corrected Spearman's R p-values used to judge statistical significance of the findings reported here.

4.3.3.4.3 Statistical analysis part 2: Effects of feature variants

4.3.3.4.3.1 Assessing the impact of signal type and filter frequency (objectives 3 and 6)

The best performing features for each primary variant (listed below) in terms of correlation strength with all clinical measures were extracted, these variants were as follows.

1) Jerk signal, mean number of peaks

2) Jerk signal, mean width
3) Acceleration signal, mean number of peaks
4) Acceleration signal, mean width

To assess the impact of filter frequency on the feature's relationship with chorea, each of the best performing features were re-computed using a sliding filter frequency. Features were generated using a 2nd order Butterworth bandpass filter with a constant low-pass cut off of 0.3Hz and a variable high-pass cut off starting at 20Hz and decreasing by 0.5Hz to a minimum of 1Hz.

The Spearman's r correlation between each feature and the clinical measure was computed, significance was assumed if $p<0.05$ and Holm-Bonferroni corrections were employed across the whole feature space.

The correlation strength of each of these features were then plotted together to visualise the difference between the signal & features types and the effect filter frequency had in terms of correlation strength.

4.3.3.4.3.2 Assessing the impact of feature task and axis composition (objectives 4 & 5)

The top performing feature across all feature variants (not including the additional filters applied in the previous subsection) were extracted. The same feature was then re-calculated using data from only the BTT and CTT in isolation, and the analysis listed in section 4.3.3.4.2 run. For the Holm-Bonferroni corrections, the BTT and CTT variants were treated as if they had been a part of the original total feature space (i.e., *n* was set to the total number of correlations run thus far plus the number of additional correlations to be run now, see the Holm-Bonferroni equation listed in section 2.3.2.5).

In total four variants of the best performing feature were extracted as follows:

1) Acceleration/jerk variant
2) x-, y-, z-axes / y-, z-axes variant
3) BTT only variant
4) CTT only variant

The correlation and regression results of each of these variants along with the original top performing feature are presented in the following results section for comparison.

4.4 Results

4.4.1 Participants

Fifty-two gene-positive participants were recruited across all studies. Table 41 shows the demographics for the cohort sub-divided by TFC stage.

Table 41: Participant demographics sub-divided by TFC stage

TFC Stage Group	n=	% Female	Age	UHDRS-TMS	Whole-body Chorea	Upper-body Chorea
Pre-manifest	4	25	33.5 (±10.1)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Prodromal	9	22.2	46.6 (±12.6)	11.8 (±6.5)	1.7 (±1.7)	0.7 (±1.0)
TFC Stage 1	16	25	55.4 (±11.0)	25.4 (±13.8)	8.8 (±4.2)	4.9 (±2.3)
TFC Stage 2	19	57.9	55.7 (±12.7)	42.5 (±14.0)	11.9 (±4.3)	7.3 (±2.7)
TFC Stage 3	4	50	46.5 (±8.7)	46.0 (±25.5)	9.8 (±7.9)	5.5 (±4.9)
Whole cohort	52	38.5	51.6 (±13.1)	28.9 (±19.6)	8.1 (±5.8)	4.7 (±3.6)

4.4.2 Results Part 1: Assessing feature relationship with chorea and the UHDRS-TMS

4.4.2.1 Histograms, scatterplots, and normality

Histograms showed varied distributions for all features. Monotonic relationships to varying degrees were present between all features and all clinical measures. Whilst some features were found to be normally distributed many were not, as such non-parametric statistical techniques were preferred. Unlike in chapter 2, no features were removed during this stage of the analysis. All histograms and scatter plots can be found in section 6.2.1.

4.4.2.2 Correlation & regression analysis

Numerous strong, significant correlations were found between each type of feature and all clinical measures. Of the 42 features initially analysed, 9 were removed after Holm-Bonferroni corrections were applied as their adjusted significance value was below 0.05. All of the removed features were the width type of feature. Notably, before Holm-Bonferroni corrections had been applied, the unadjusted p-values of all but one of the removed features were already above 0.05. Thus, it is likely these features are not sensitive to chorea in HD. All correlation and regression analysis results can be found in section 6.2.2 Table 50, and Table 51.

4.4.3 Results Part 2: Effects of feature variants

4.4.3.1 Assessing the impact of signal type and filter frequency (objectives 3 and 6)

The best performing features out of each feature type is shown below in Table 42 and their associated scatter plots against each of the clinical measures in Figure 38.

As shown in Table 42, jerk features outperformed acceleration features both in terms of their relationship with chorea and the UHDRS-TMS. In both acceleration and jerk signal, peak type features outperformed width type features.

Table 43 and Figure 38 show how the correlation strength changes for these features as they are generated using different filter frequencies. As can be seen, filter frequency appeared to have a moderate impact on correlation strength.

The best overall feature was the number of peaks counted along the y- and z-axes of the jerk signal using high pass cut-off of 7.5Hz.

*Table 42: Performance of top features for each variant – jerk peaks, jerk widths, acceleration peaks, acceleration. *** indicates p<0.001*

Whole-body Chorea				
Feature		Spearman's R	MAE (± std)	Normalised MAE (%)
Top Jerk Peaks	Jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.81***	2.9(±0.6)	15.3
Top Jerk Widths	Jerk signal widths (x-, y-, z-axes; 3Hz filter)	-0.72***	3.5(±0.6)	18.5
Top Acceleration Peaks	Acceleration signal peaks (x-, y-, z-axes, 3Hz filter)	0.71***	3.5(±0.9)	18.6
Top Acceleration Widths	Acceleration signal widths (x-, y-, z-axes; 3Hz filter)	-0.64***	3.8(±0.7)	20.3
Upper-body Chorea				
Feature		Spearman's R	MAE (± std)	Normalised MAE (%)
Top Jerk Peaks	Jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.79***	1.8(±0.4)	14.8
Top Jerk Widths	Jerk signal widths (x-, y-, z-axes; 3Hz filter)	-0.72***	2.2(±0.4)	18.0

Top Acceleration Peaks	Acceleration signal peaks (x-, y-, z-axes, 3Hz filter)	0.71***	2.1(±0.4)	17.7
Top Acceleration Widths	Acceleration signal widths (x-, y-, z-axes; 3Hz filter)	-0.64***	2.3(±0.4)	19.5
UHDRS-TMS				
Feature		Spearman's R	MAE (± std)	Normalised MAE (%)
Top Jerk Peaks	Top Jerk Peaks: Jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.85***	9.5(±2.3)	12.2
Top Jerk Widths	Jerk signal widths (x-, y-, z-axes; 3Hz filter)	-0.77***	11.1(±3.0)	14.3
Top Acceleration Peaks	Acceleration signal peaks (x-, y-, z-axes, 3Hz filter)	0.79***	10.6(±2.4)	13.6
Top Acceleration Widths	Acceleration signal widths (x-, y-, z-axes; 3Hz filter)	-0.64***	12.2(±3.0)	15.7

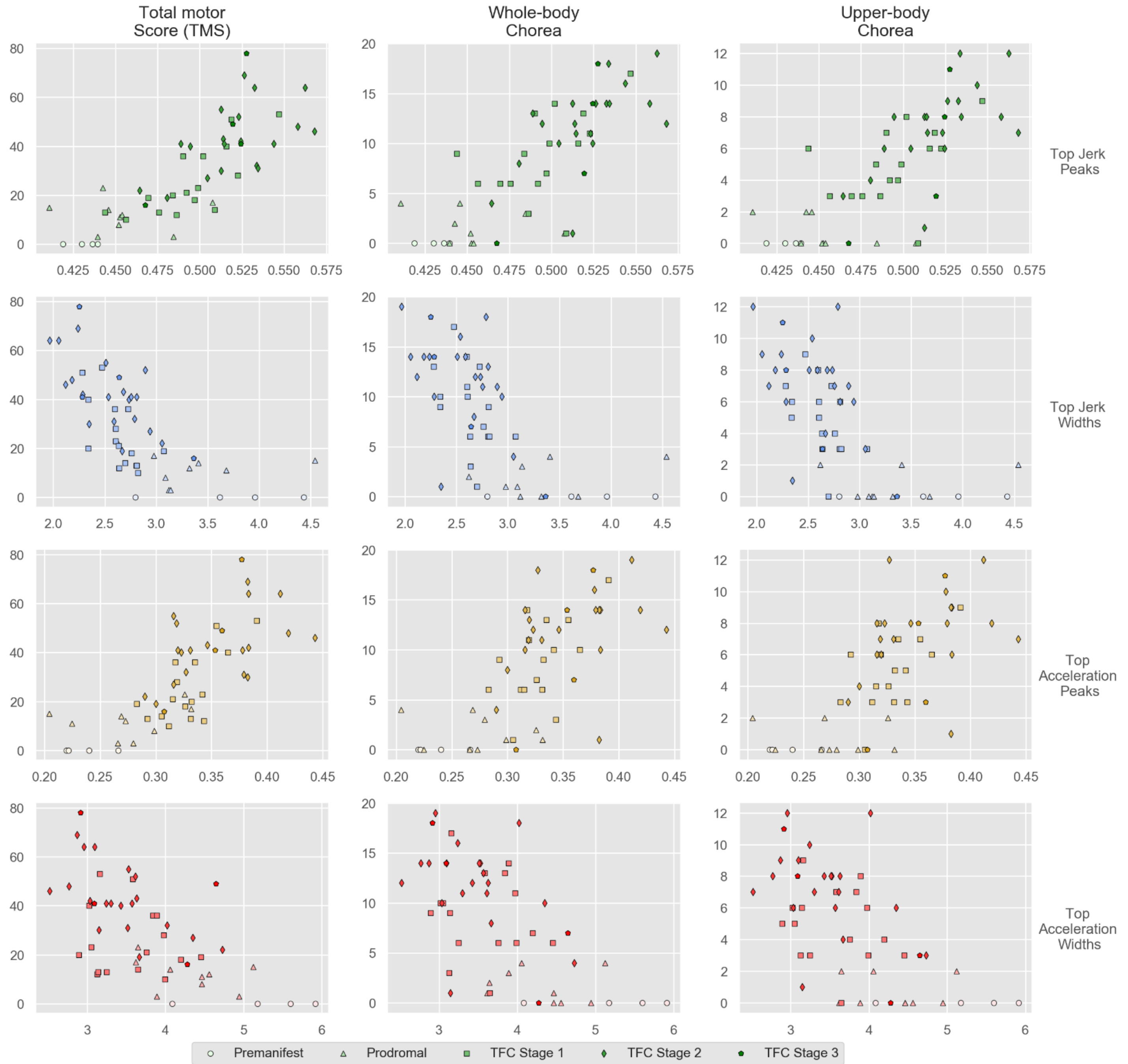


Figure 38: Scatter plots of the top features and UHDRS-TMS (left column), whole-body chorea (middle column) and upper-body chorea (right column). The top jerk peak feature is shown on the top row, the top jerk width feature on the 2nd row, the top acceleration peak feature on the 3rd row, and the top acceleration width feature on the fourth row. Feature values are shown on the x-axis, clinical measure values are shown on the y-axis.

Table 43: Spearman's R correlation statistics for each of the best performing features shown in Table 42 across different filter frequencies for each clinical measure. Greyed out boxes indicate no statistical significance found after Holm-Bonferroni corrections. The top performing feature for each feature type is surrounded by a thicker line and shown in bold.

Jerk signal peaks (y-, z-axes)			
High-pass Filter	UHDRS-TMS	Whole-body Chorea	Upper-body Chorea
3Hz	0.83	0.75	0.74
7.5Hz	**0.85**	**0.81**	**0.79**
13Hz	0.81	0.76	0.74
20Hz	0.79	0.74	0.71
Jerk signal widths (x-, y-, z-axes)			
High-pass Filter	UHDRS-TMS	Whole-body Chorea	Upper-body Chorea
3Hz	**-0.77**	**-0.72**	**-0.72**
7.5Hz	-0.51	-0.50	-0.50
13Hz	-0.44		
20Hz			
Acceleration signal peaks (x-, y-, z-axes)			
High-pass Filter	UHDRS-TMS	Whole-body Chorea	Upper-body Chorea
3Hz	**0.79**	**0.71**	**0.71**
7.5Hz	0.65	0.64	0.63
13Hz	0.61	0.62	0.61
20Hz	0.60	0.61	0.60
Acceleration signal widths (x-, y-, z-axes)			
High-pass Filter	UHDRS-TMS	Whole-body Chorea	Upper-body Chorea
3Hz	**-0.64**	**-0.64**	**-0.64**
7.5Hz	-0.45	-0.47	-0.47
13Hz		-0.45	-0.46
20Hz			-0.44

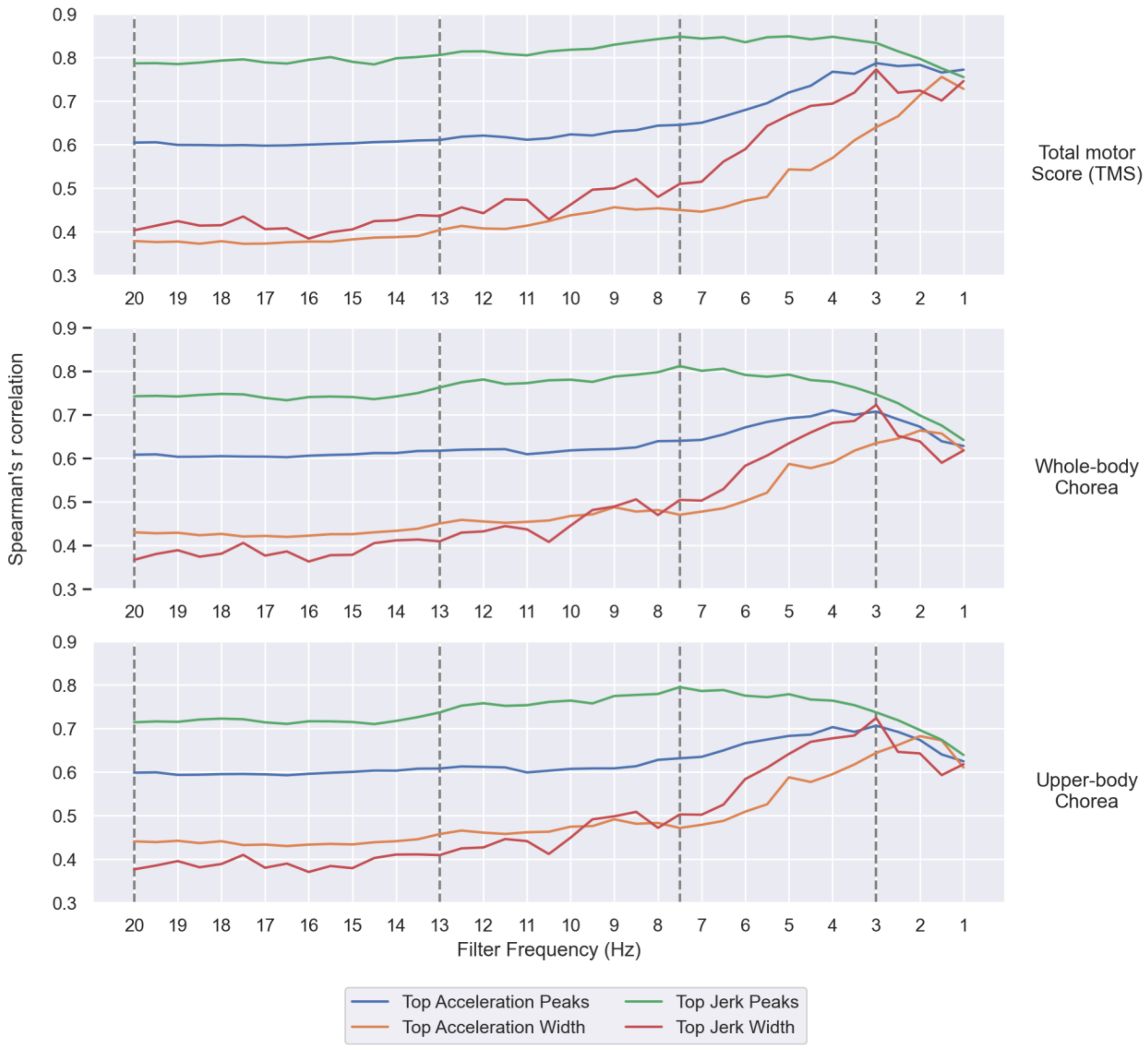


Figure 39: Line graphs for each of the clinical measures showing how performance of the top features changes as they are filtered using different high pass cut-offs. The y-axis is the Spearman's R correlation. The x-axis is the filter frequency used. Grey dashed lines indicate the four filter frequencies that were initially selected for analysis based on the spectral edge frequencies.

4.4.3.2 Assessing the impact of feature task and axis composition (objectives 4 & 5)

The best overall feature in terms of correlation strength with the clinical measures was found to be the number of peaks counted along the y- and z-axes of the jerk signal using high pass cut-off of 7.5Hz. Table 44 shows the performance of three variants of this feature as follows.

- Alternative axis variant (x-, y-, z-axes)
- BTT only

- CTT only

Although the optimal feature was calculated using just the y- and z-axes from both tasks the difference between the variants is minimal

*Table 44: Performance of different variants of best overall feature. The best overall feature is bolded and shown surrounded by bold lines. Variants are the same feature generated in different ways – (1) using the x-, y-, z-axes, (2) using just the BTT task, and (3) using just the CTT task. *** indicates p<0.001*

Whole-body Chorea			
Feature	Spearman's R	MAE (± std)	Normalised MAE (%)
Jerk signal peaks (y-, z-axes, 7.5Hz filter)	**0.81*****	**2.9(±0.6)**	**15.3**
Jerk signal peaks (**x-, y-, z-axes**, 7.5Hz filter)	0.77***	3.2(±0.7)	16.6
BTT jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.76***	3.1(±0.7)	16.2
CTT jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.76***	3.2(±0.7)	16.9
Upper-body Chorea			
Feature	Spearman's R	MAE (± std)	Normalised MAE (%)
Jerk signal peaks (y-, z-axes, 7.5Hz filter)	**0.79*****	**1.8(±0.4)**	**14.8**
Jerk signal peaks (**x-, y-, z-axes**, 7.5Hz filter)	0.75***	1.9(±0.5)	16.2
BTT jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.74***	1.9(±0.4)	15.5
CTT jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.74***	2.0(±0.6)	16.3
Total Motor Score (UHDRS-TMS)			
Feature	Spearman's R	MAE (± std)	Normalised MAE (%)
Jerk signal peaks (y-, z-axes, 7.5Hz filter)	**0.85*****	**9.5(±2.3)**	**12.2**
Jerk signal peaks (**x-, y-, z-axes**, 7.5Hz filter)	0.84***	9.5(±2.2)	12.2
BTT jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.84***	9.1(±2.4)	11.7
CTT jerk signal peaks (y-, z-axes, 7.5Hz filter)	0.78***	10.8(±2.5)	13.9

4.5 Discussion

Due to the large number of chapter objectives defined in section 4.2, this section is split up by objective with an overall conclusion of the study presented in section 4.7.

4.5.1 Objective 1: Assess the relationship between signal features generated during the instrumented C3t with clinical measures of chorea and the UHDRS-TMS

This study provides evidence that an instrumented version of the C3t can be used to produce features from wrist-worn accelerometers that are both highly correlated with and can estimate with a high degree of accuracy both chorea and the UHDRS-TMS.

Two types of features were generated – the number of peaks and the mean width between those peaks. Statistically significant, positive monotonic relationships were found between the number of peaks and the clinical measures whilst negative monotonic relationships were found between the mean width between peaks and the clinical measures. The number of peaks outperformed the mean width features for all clinical measures, suggesting that the peaks regularity in the signal is more relevant to chorea than the width between peaks.

In contrast to the lack of relationship observed between the non-instrumented C3t time scores and chorea shown in chapter 2, the extracted features have a similar level of correlation with chorea as the C3t time scores do with UHDRS-TMS. This illustrates the point made in chapter 1 – that specific, tailored sensor features can be very powerful for finding relationships with motor symptoms relative to non-tailored traditional ones (e.g., task time scores).

The difference in performance is likely because, as was discussed in chapter 2, that the C3t time scores will be impacted by almost all motor symptoms seen in HD as well as many non-motor symptoms. In contrast the sensor features used study were extracted specifically because they were thought likely to be sensitive to chorea.

As was discussed in section 4.2, the UHDRS-TMS is comprised of multiple motor symptoms and so an instrumented assessment that is thought to be sensitive to the UHDRS-TMS must, by definition, be sensitive to all or at least most of its component symptoms. The relationship between the extracted features and the UHDRS-TMS however is will not be due to the developed features being sensitive to all motor symptoms in HD. On the other hand, the high weighting of chorea in the UHDRS-TMS, coupled with the relatively early stage the studied cohort (~55% TFC Stage 1 or earlier, ~92% TFC Stage 2 or earlier) and the dominating presence of chorea during these early stages suggests a clear clinical rationale for this finding.

As in chapter 2 it was found that the non-instrumented C3t time scores are unrelated to chorea, the next most relevant assessment to compare the instrumented C3t to is choreomotography. As discussed in 1.4.3.2, choreomotography has the participant grip and attempt to hold stationary a weighted object equipped with IMUs for 35 seconds. Two features are produced from the resultant signal - the change in position and change in rotation the object. Reilmann *et al.*, (2011) reported a moderate correlation between their choreomotography assessment and whole-body chorea (r=0.48) along with a strong correlation with the UHDRS-TMS (r=0.64).

Significantly stronger relationships were found in this study between some of the extracted features and both whole-body chorea (r=0.81) and the UHDRS-TMS (r=0.85) than were for choreomotography by Reilmann *et al.*, (2011). The reason for this may be due to choreomotography using change in position as its feature. As was pointed out by Casula *et al.*, (2018), changes in position do not necessarily relate to just chorea, but all involuntary movement. Of course, likewise, it is reasonable to suggest that the peaks in acceleration/jerk signals could be caused by any involuntary movement. However, the analysis presented here does suggest that the number of peaks in a jerk signal during the C3t is a more sensitive feature for chorea than the sum of changes in position used in choreomotography.

Similar to Bennasar *et al.*, (2018a) the extracted features were further assessed by attempting to estimate the clinical measures. The best performing feature was found to be capable of accurately estimating whole-body chorea (84.7% accuracy, 15.3% error), upper-body chorea (85.2% accuracy, 14.8% error) and the UHDRS-TMS (87.8% accuracy, 12.2% error).

In summary this chapter has shown that features can be extracted from the instrumented C3t which are highly correlated with chorea. As discussed in section 1.5.2.4, this is the first step required to show evidence of new assessments validity. Moving forwards, the next step should be to confirm whether the features identified here are indeed sensitive to chorea (e.g., by using expert-labelled video reference data), and to adjust the methodology applied here to enhance that sensitivity if possible (e.g., by seeing whether the sensitivity increases when fixed-length timing windows are used). These next steps are unfortunately beyond the scope of this thesis and are discussed in sections 4.6 and 5.3. Additionally, as noted in section 4.3.2.1, it was assumed that the inclusion of participants from different disease stages would not affect the results of this study outside of the required impact of different levels of chorea severity. To obtain participants with different levels of chorea symptoms recruiting participants from across the disease stage spectrum was required. However, whether there is any relationship between the developed features and disease stages should be assessed, as is recommended in section 4.6.7.3.

4.5.2 Objective 2: To use features for objective 1 that are simple to translate into clinical practice

Whilst the primary objective of this study was to develop features linked to chorea, a core secondary objective was that the relationship made sense clinically. The reason for this is primarily to aid clinical uptake – it is much easier to trust a finding when it makes intuitive sense.

Two types of features were used in this study - the total number of peaks and the mean width between the peaks. These features were chosen based on the clinical description of chorea and its progression.

Choreatic movements produce random, sudden jerk-like movements (Jankovic and Roos, 2014). As chorea worsens, frequency and severity of these movements will increase (Kieburtz *et al.*, 1996). A sudden change in acceleration, which mechanically is what chorea does, will produce a spike in both the acceleration signals and its time derivative, jerk. As such, one could reasonably expect that the more choreatic movements that are made, the more of these spikes/peaks would be seen in the recorded signals and the distance between them would decrease (i.e., they occur more overall and more often).

The results presented in this study support the rationale of the chosen features. Strong *positive* correlations were observed between the total number of peaks in the signals (both acceleration and jerk) and measures of chorea. Similarly, strong *negative* correlations were observed between the mean width between peaks (again in both acceleration and jerk signals) and measures of chorea.

In clinical terms, these results can be explained thusly - as the severity of chorea increases in a patient, the number of peaks they have in their acceleration/jerk signals increases and the distance between those peaks decreases.

As the sensors are only worn on the hands it could be questioned why there would be a relationship with whole-body chorea and the upper-body chorea components that are not the left/right upper extremities (arms). The reasoning is the same as that used by Reilmann *et al.*, (2011), namely that despite the sensor being attached to only one specific area any choreatic movement throughout the body will likely cause a sudden alteration in acceleration. If for example the torso jerks forward during the C3t, this will naturally have a knock-on effect on the arms placed on the sensors. This is an aspect of the instrumented C3t which needs investigation. As such, a clear limitation of this study is the absence of accelerometers attached to the participants torso, as is discussed in section 4.6.

It should be noted that whilst the peaks in the signals can be explained by chorea, not every peak is likely due to a choreatic movement. Other hyperkinetic disturbances would logically cause similar peaks. Similarly, ordinary movements during the C3t could cause peaks in the signal to occur. Although there appears to be a link between number and distance between peaks in the signals and the UHDRS motor assessment chorea sub-items, it is impossible to know which, if any, peaks were due to choreatic movements without video reference data. This is also clear limitation of this study and so will be discussed in section 4.6.

4.5.3 Objective 3: To assess whether features generated from jerk were better for assessing chorea than identical features generated from acceleration

At the start of this study, it was unknown whether features calculated from the jerk signal would be empirically better than the same features generated from acceleration. Whilst acceleration is commonly utilised its derivative jerk is arguably less well known although it is has been used in previous studies of movement disorders (Park, 2016; Lapinski *et al.*, 2019; Teshuva *et al.*, 2019; Zhang *et al.*, 2019). As was argued in section 4.3.3.3.2, it was assumed that the sudden changes in acceleration caused by choreatic movements would be more noticeable in accelerations derivative, jerk, than in acceleration itself.

The difference observed between the best performing jerk feature and best performing acceleration feature was not substantial for any of the clinical measures. The best performing jerk feature showed a 0.1, 0.08, and 0.06 increase in correlation strength and 3.3%, 2.9%, and 1.4% increase in estimation accuracy relative to the best performing acceleration feature for whole-body chorea, and upper-body chorea and the UHDRS-TMS respectively.

These results suggest that whilst there may be some benefit in using jerk over acceleration for assessing chorea, the type of feature used (i.e., peaks or widths) and the applied filter frequency had a much larger effect. Similar future work may wish however to include features generated from jerk signals in their feature set, as there does seem to be some increase in correlation strength and estimation accuracy relative to comparable acceleration features.

4.5.4 Objective 4: To assess whether the inclusion/exclusion of the x-axis from generated features had an impact on the feature's relationship with chorea

A key aspect of the C3t is that the voluntary/intentional movement is primarily contained along the horizontal plane of the board. This corresponds to the x-axis of the wrist-worn accelerometers used in this study. As the goal of this study was to assess chorea, which is characterised by *involuntary* movements, it was hypothesised that the accuracy of the generated features might increase if the x-axes were excluded, as this would in theory remove the majority of the voluntary/intentional movements. To test this, two sets of features were generated, one which included the x-axis and one which excluded it.

The best-performing feature came from the set that excluded the x-axis, however the difference between it and the same feature which included the x-axis was minor (0.04, 0.04, 0.01 increase in correlation strength increase and 1.3%, 1.4%, 0% estimation accuracy increase for whole-body chorea, upper-body chorea, and UHDRS-TMS respectively). This difference is so small it is plausibly due to random chance alone. As such, it can only be concluded that in this study including/excluding the x-axes has limited practical difference on feature performance. The effect of including and excluding the x-axis was however important to take into account in case the observed effect had been large. Future work may wish to take similar considerations into account.

4.5.5 Objective 5: To assess whether the features generated from both the BTT and CTT were significantly superior to features generated from just the BTT or CTT

An integral part of the C3t is the various different tasks it contains. As was discussed at length in chapter 2, in a clinical setting the numerous tasks the C3t contains could hinder its uptake due to the time requirements of performing and setting up each task. As such in this study the performance difference (in terms of correlation strength & estimation accuracy of the clinical scores) between the best performing feature generated using both of the studied tasks (i.e., dual task features), and the same feature generated using one of the studied tasks at a time (i.e., single task features) were compared.

The correlation strength and estimation accuracy of the best performing dual task feature slightly outperformed the single task features. The largest observed difference was a 0.05, 0.05, and 0.07 increase in correlation strength and a 1.6%, 1.5%, and 1.7% increase in accuracy for whole-body chorea, upper-body chorea, and the UHDRS-TMS respectively. The effect appears to be larger than the effect observed by including/excluding the x-axis, but again the difference is so small it may be just due to random chance. Notably, aside from the estimation accuracy of the UHDRS-TMS there was almost no difference in correlation strength or estimation accuracy between the single task features (0.00, 0.00, 0.06 correlation strength difference, 0.7%, 0.8%, 2.2% estimation accuracy difference for whole-body chorea, upper-body chorea, and the UHDRS-TMS respectively).

The small observed increase could be due to the random nature of chorea as a movement disorder. Choreatic movements occur (seemingly) randomly and so, as with any randomly occurring phenomena, the longer it is observed the more accurate a 'picture' of it can be obtained. Parallel to this, the features used in this study are based around the number peaks that occur in the recorded signals. As choreatic movements are random the longer a recorded signal is the more chance a choreatic movement has to occur at any given point in the signal. As was discussed in chapter 2, the increase in the C3t task times could be attributed to multiple factors not just chorea. As such, in this study we time-normalised the recorded signals in order to make the structure of them more comparable. However, the dual task features still ultimately contain double the amount of data points as the single task features, which may explain why they provide a better ‘snapshot’ of chorea than the single task features, even if only marginally so.

The similarity of the results seen between the single task variants is like the results reported in chapter 2. This suggests again that the BTT and CTT are similar enough to each other such that symptom severity levels impact task performance equally. It has been suggested that an increased cognitive load will impact motor performance (Fritz *et al.*, 2016). Based on this rationale, the CTT was developed with a cognitive load that was supposed to be notably higher than that of the BTT. However, the similarity in performance (in terms of correlation strength and estimation accuracy for accepted HD clinical assessments) between the BTT and CTT seen now in both this study as well as in chapter 2 suggests this load may be insufficient. It is however important to consider that participants across a range of disease stages were recruited for this study and were analysed together. It could conceivably be the case that participants at later disease stages are impacted by the increased cognitive load present in the CTT. Studying the differences between groups with different levels of cognitive load (and thus likely at the higher end of the disease stage spectrum) was however not the purpose of this study, and so is recommended as a direction for future work in section 4.6.7.3.

Overall, it seems reasonable to suggest that if the absolute highest estimation accuracy of chorea is required then both the BTT and CTT are required. It should be noted that unfortunately, as was the case in chapter 2, the DTT was not included in the analysis. This is a limitation of this study which can be used to inform future work as will be discussed in section 4.6.

4.5.6 Objective 6: To assess what impact filter frequency had on the feature’s relationship with the clinical measures

Unlike other movement disorders, such as bradykinesia and tremor in PD, the optimal frequency band for assessing chorea is unknown. As the frequency band of a signal can have considerable impact on features generated from that signal, identifying an appropriate filter band is important for developing high quality features. A heuristic approach was taken here to understand the effect of different frequency filters, by first generating features using four high pass cut offs bands – 20Hz, 13Hz, 7.5Hz and 3Hz. The lower three cut offs were decided calculating the mean 95th, 90th and 75th spectral edge frequencies of all participants’ signals. Post-hoc, the best performing feature for each signal type was re-generated using a sliding set of high pass filters, starting at 20Hz, and working down to 1Hz with a 0.5Hz difference between each filter, to assess the difference more fully.

As has been shown in this study, the high-pass filter frequency selected had varying degrees of impact on the strength of the relationship between the features and clinical measures. The effect of different filter frequencies was particularly pronounced in the width features. The jerk width feature was for example found to not be statistically significantly associated with whole-body chorea using a 20Hz and 13Hz filter but was using a 7.5Hz and 3Hz filter where the strength was -0.5 and -0.72 respectively. This effect was least pronounced with the best overall feature, the jerk signal peaks, where all features were found statistically significant, and the biggest performance gain was for upper-body chorea (0.08 difference between 20Hz (r=0.71) and 7.5Hz (r=0.79)).

These results highlight the importance of filtering acceleration data before generating features from it and suggest that chorea may be best observed by removing frequencies higher than 7.5Hz. also appears to show a steady increase in correlation strength for the best performing feature as the high pass cut off is reduced, until at very low frequency cut offs the strength starts to rapidly drop.

Further work is required to fully understand the generated features and their relationship with chorea, in particular video reference data is required to understand which peaks are true choreatic movements. However, future work may wish to consider bandpass filtering acceleration signals with a high-pass filter of 7.5Hz and a low-pass filter of 0.3Hz as this frequency band was found to be optimal for generating features sensitive to chorea in this study.

It is also notable that the method used to identify promising filter frequencies, looking at the mean of the 95th 90th and 75th spectral edge frequencies, produced good estimates of the optimal filter frequency. Whilst this is a heuristic method it was found to be highly effective in this study. As shown by the dashed lines in Figure 39 which indicates the 4 filter frequencies chosen, the filters selected in this manner produced the optimal features for all but one feature type. As such, when the optimal frequency filter is unknown, this method could be used as an automated heuristic approach for filtering signals before features are extracted.

4.6 Limitations and future work

4.6.1 Limitations Overview

The following subsections detail each identified limitation of this study, explain why the limitation matters, and propose future work directions to address them. In the subsection an additional suggestion for future work not based on the limitations of this study is suggested.

4.6.2 Limitation 1: Lack of video reference data

Video reference data (i.e., video recordings of participants performing the instrumented C3t) is needed to fully understand the results of this study. The primary finding of this study is that the mean number of peaks, and the distance between them, are both strongly related to whole-body and upper-body chorea in HD. However, as was discussed in section 4.5.2, without video reference data we cannot be certain which (if any) of the peaks are due to true choreatic movements. Notably, the criticism of choreomotography by Casula *et al.*, (2018) feeling the assessment would be affected by any involuntary movement, not just chorea. This is equally the case for the features used here. Whilst the results of this study show stronger relationships with chorea than the work on choreomotography by Reilmann *et al.*, (2011), video reference data is needed substantiate any claim we may make that the instrumented C3t can be used to assess chorea.

Future work can address this limitation by filming participants whilst they take the instrumented C3t. The peaks in the acceleration/jerk signals can then be labelled and tied to corresponding time points in the video reference data. Expert clinicians can then observe the movements around these time points, and label them as either a choreatic movement, another type of involuntary movement, or as a purposeful movement. This would allow us to better understand the relationship of the detected peaks and choreatic movements by treating the instrumented C3t as a rater and assessing the inter-rater reliability between it and the expert clinicians.

Importantly it should be remembered that in this study the signals were time-normalised to reduce the impact of different participants taking different lengths of time for reasons other than chorea. However, if peaks in the signals are to be tied to specific time points in video reference data then signal cannot be time normalised. Thus, such a study would need to use a modified version of the instrumented C3t where participants transfer as many tokens as possible in a set amount of time.

4.6.3 Limitation 2: Lack of longitudinal data

This study was based entirely on cross-sectional instrumented C3t data which, whilst sufficient for studying the relationship between it and chorea, is insufficient for studying its relationship with the *progression* of chorea.

As has been discussed throughout this thesis, there is a need for assessments in HD sensitive to early-stage motor symptom progression. One of the rationales for developing an instrumented assessment of chorea is that chorea is often one of the first motor symptoms to present in HD making it an ideal symptom for helping assess the progression of early-stage HD. As such, whilst this study is a useful first step in realising the instrumented C3t as a sensitive assessment for chorea, the next step is determining whether it is also sensitive to the evolution of chorea in a patient over time.

Similarly, the lack of longitudinal data means that we cannot assess any test-retest effect that may be present in the instrumented C3t. As was discussed in chapter 2, assessing whether an assessment has a test-retest effect (i.e., whether participants who have performed the test multiple times have different performance results than participants who have not taken the test before) is vital for showing a tests validity.

Future work can address both of these limitations by collecting longitudinal data. Instrumented C3t data will need to be collected at regular intervals over a short time period to properly assess test-retest effects. Similarly, data will need to be collected (along with clinical measures) over a long period of time to assess whether the produced measures are sensitive to the change in chorea over time. As the measures captured during the non-instrumented C3t can still be recorded during the instrumented C3t such work could also address the need for longitudinal data of the non-instrumented C3t as discussed in chapter 2.

4.6.4 Limitation 3: Small pre-manifest and prodromal participant sample size

The total sample size used in this study was 52. Whilst this sample size is comparable to other studies of a similar nature (Reilmann *et al.*, 2011a; Bennasar *et al.*, 2018), a larger sample size would bring with it a higher degree of confidence, although this can be said for most studies. More importantly there is a lack of pre-manifest and prodromal participants in the studied cohort (n=4; n=9 respectively).

As has been stated, one of the rationales for focusing on generating features from the instrumented C3t thought to be sensitive to chorea was chorea's prominence during these earlier stages and the UHDRS motor assessment's low sensitivity to progression during these stages. Whilst noticing progression would require longitudinal data (the lack of which is a limitation in its own right as previously discussed), had more pre-manifest and prodromal data been collected a clearer picture could have been developed as to whether chorea during these early stages is detectable by the instrumented C3t.

Future work should look to add additional pre-manifest and prodromal participants to the available instrumented C3t dataset. Longitudinal data of this cohort would be particularly interesting as it would allow the investigation of whether the instrumented C3t can detect chorea emerging as participants progress through the pre-manifest, prodromal, and early manifest stages of HD.

4.6.5 Limitation 4: Lack of Dual Transfer Task analysis

The C3t contains 6 tasks 3 of which are transfer tasks and are so suitable for analysis using accelerometers. In this study only the BTT and CTT were studied. The reason for this was, as was the case in chapter 2, significantly more data were available for the BTT and CTT relative to the DTT as these tasks were present in both the C3t and MBT.

However, as was mentioned in the results, little difference in terms of feature performance was observed between the BTT and CTT despite it being theorised that the addition of a cognitive load should result in greater movement disorder severity. The DTT however contains a much greater cognitive load than the CTT and so may elicit the increase in motor symptoms, and so affect test performance, in the manner it was thought the CTT would. As the DTT was not included in this study, whether the cognitive load of the DTT is sufficient to elicit such an increase in motor symptoms is unknown.

Future work should seek to increase the amount of data available to the DTT such that the analysis presented in this study can be conducted and the usefulness of cognitive load of the DTT evaluated.

4.6.6 Limitation 5: No assessment of the impact down sampling the acceleration signal

In this study four pre-processing steps were applied to the captured signals in the following order – segmentation of the signal into specific tasks, filtering of the signal using a range of frequency cut-offs, conversion into jerk, and normalisation/down sampling (see section 4.3.3.2). The impact all but one of these steps had on the results of the correlation & regression analysis were, to an extent, explored in this study. However, the impact down sampling had on the features generated was not explored.

The underlying rationale for down-sampling / normalising the captured signals was to reduce the impact of task times on the signal, as the underlying task participants were performing was the same (see section 4.3.3.2.5). The signals themselves are naturally not flat, and the longer the participants take to perform a task the more peaks & troughs will be recorded, especially given the raw sampling rate of 100Hz which is naturally quite noisy. As the two features generated are dependent on the number of peaks present in the signal, we wanted to reduce the impact of the time the task took as much as possible, removing noise from the signal and attempting to retain only the 'true' peaks & troughs. The impact of this process is shown in Figure 36. It should be noted that similar approaches are taken in gait analysis, where the 'shape' of the resultant signal is what matters, rather than the time taken to complete a gait cycle (Whittle, 2007).

However, the lack of investigation into the impact of down sampling the signals should have been investigated in this study and thus should be considered a limitation of this work. It is feasible that alternative down sampling strategies might be preferable, or even conversely that down sampling has a negligible effect on the produced features (although we consider this to be unlikely). Regardless of whatever impact down sampling is found to have, the fact remains that it should be investigated in order to provide firm recommendations about whether it is necessary or not. As such, future work should seek to explore the impact down sampling has on the produced features relative to the raw signals (and potentially other degrees of down sampling).

4.6.7 Directions for additional future work

4.6.7.1 Assessing additional HD movement disorders & exploring additional features

Future work should seek to address the limitations previously listed to confirm the findings presented in this study. However, despite its limitations this study does present evidence that features can be generated from the instrumented C3t that are related to chorea in HD. As such, other similar work should be conducted for other motor symptoms seen in HD. How future work might approach doing so for four particular motor disorders, namely bradykinesia, dystonia, rigidity, and oculomotor dysfunction, is briefly discussed below.

Bradykinesia, defined as reduced movement velocity/slowness to initiate movement, has been previously detected and assessed using instrumented assessments in PD (Rovini, Maremmani and Cavallo, 2017). Measures such as the mean, min, and max velocity of participants during the instrumented C3t may be a simple feature sensitive to bradykinesia. Measures of a participant's slowness to initiate could be come from looking at the speed at which participants switch from one 'stage' of a task to another (e.g., time taken to start moving after the task starts, mean time to switch between pickup and token transfer stages). These two suggested features would likely require IMUs (in order to calculate velocity) and video reference data (in order to determine the timestamp participants started doing each movement stage), respectively.

Dystonia is defined as repetitive, twisting movements resulting in abnormal fixed postures. The as 'twisting' is a key aspect of dystonia it is likely gyroscopic information from IMUs would be needed to generate features sensitive to it. A single IMU could be used to detect abnormal & repetitive 'twisting' movements by looking at the amount of rotation of the sensor during tasks. A series of IMUs along the arms and torso might allow for 'abnormal postures' to be detected during the tasks by calculating joint angles.

Rigidity is defined as stiff, inflexible muscles. IMUs could again be used to calculate the variation of joint angles which one might expect to be lower in the presence of stiff/inflexible muscles.

Oculomotor dysfunction presents in HD as delayed/suppressed saccade initiations and gaze impersistence (i.e., difficulty tracking moving objects and difficulty maintaining focus on a given object). The movement of the C3t tokens across the board presents an opportunity to embed sensors into the board itself suitable for conducting eye movement recordings. Such recordings have been previously shown to be useful as quantitative, objective measures of oculomotor suitable for monitoring disease progression (Clark *et al.*, 2019).

Although additional instruments may be needed to explore some other movement disorders, there is still now a good deal of instrumented C3t data to work with. As was mentioned in section 4.3.3.3.2, there are a great many features which could be potentially generated from the data collected for this study. Whilst this study focused only on two features (in order to, among several reasons, enhance clinical translation) other features which could be generated from accelerometer & C3t may give insights into HD as well as the C3t. Future work could, for example, seek to simplify the method presented by Bennasar *et al.*, (2018) and use instead a small selection of features in an attempt to classify between control and HD populations. As another example, future work may wish to explore the difference in acceleration signals between the different C3t tasks. A measure of entropy for example, might give insights into whether participants generate typically more complex signals as the tasks progress (e.g., from mental fatigue, or the increased cognitive load hypothesised to be present in later transfer tasks). Regardless of either of these specific suggestions is followed up, future work should seek to take advantage of the corpus of instrumented C3t data currently available and explore the myriad features of acceleration signals which could give further insight into HD.

4.6.7.2 Assessing identified features in control populations

As was stated in section 4.3.2.1, control data was not used in the study. This was because UHDRS motor assessment scores were not available for control participants with instrumented C3t data, and this study was focused on whether features generated from the instrumented C3t could be used to estimate the severity of chorea, not whether they could distinguish between control and HD populations. It should be noted that previous work has already shown that sensor features extracted during the C3t can be used to distinguish with a high degree of accuracy between HD and control populations (Bennasar *et al.*, 2018).

Future work should seek to understand how the identified features presented in this study behave in control populations. This is particularly important given that this study shows a strong relationship between the instrumented C3t and chorea. As was discussed in section 1.3.3, pre-manifest and prodromal HD can be difficult to detect using traditional methods and chorea is often one of the first motor symptoms seen throughout these stages. As such, the features found in this study to be sensitive to chorea may be sensitive to early-stage changes in pre-manifest and prodromal HD to a degree which allows them to be distinguished from control populations.

4.6.7.3 Assessing identified features across the disease stage spectrum & cognitive loads

As was noted in section 4.3.2.1, the studied sample consisted of individuals from across the HD disease stage spectrum, ranging from pre-manifest, to prodromal, to TFC Stages 1-3. In this study the inclusion of participants from across the disease stage spectrum was felt to be prudent, as it is effectively an unavoidable by-product of wanting to study participants with various degrees of chorea in HD. However, the effect of including participants at different disease stages on the developed features should be investigated, as the developed features could capture important information relating to disease stage.

Similarly, whether participants with greater degrees of cognitive dysfunction where more heavily impacted by the increased cognitive load of the CTT relative to the BTT in terms of the developed features was not explored in this study. This is important to consider as if the CTT does in fact increase motor symptoms seen in participants with greater degrees of cognitive dysfunction this effect will need to be taken into account in subsequent models, and as such future work should seek to explore this.

4.6.7.4 Assessing the identified features in a fixed time C3t task

An interesting direction for future development would be to see how the developed features perform in a fixed time variant of a C3t transfer task. As was discussed in section 4.3.3, chorea occurs with seemingly random frequency and is may there be better picked up by looking at signals recorded over long time periods (as over a longer period there is more likelihood of choreatic movements occurring and so estimating how regular they are will become more accurate). Unfortunately recording over long time periods (especially continuous recording in the home) is naturally more resource intensive than recording in short bursts during a clinical assessment like the C3t and would be difficult to justify without prior evidence that the approach would be feasible.

However, because of this study we now know that accelerometers used during the C3t can be used to generate features highly correlated with chorea. If one wants to move towards continuous in-home monitoring, a natural next step prior to doing so would be to simulate conditions closer to continuous monitoring producing a fixed-time (rather than time-variable) C3t task. Such a task could provide the participant with a large number of tokens and ask them to transfer as tokens into a container over a fixed period of time, making it clear the number of tokens transferred doesn't matter. If the developed features were still highly correlated with chorea, then this would provide further evidence that continuous in-home monitoring might be feasible. Whilst in-home monitoring has not been discussed in this thesis, as it is thoroughly out of scope, it is important to note that it has been highly effective in PD and other chronic conditions (Patel *et al.*, 2012; Pulliam *et al.*, 2014).

In my opinion, continuous in-home monitoring of HD using sensors is the natural endpoint for monitoring the diseases evolution. The ultimate goal of assessing HD motor symptoms using sensors is to be able to detect small changes over time in response to treatment. If we can monitor symptoms evolve over time in the home, then we should be able to apply the same technology to do the reverse, in response to treatment.

4.7 Conclusions

There are six conclusions which may be drawn from this study.

First, this study has shown that accelerometers worn whilst taking the instrumented C3t can be used to produce features that are significantly correlated with whole-body and upper-body chorea. These same measures can also be used to estimate whole-body and upper-body chorea with a reasonable degree of accuracy. The observed relationship between chorea in HD and the instrumented assessment presented here is stronger than that of the most relevant previous study we are aware of (Reilmann *et al.*, 2011b). There may be some value in generating features from jerk signals rather than acceleration signals, however the main driver of performance appears to be in the selection of features and the frequencies used to filter the signal. As has been discussed, the differences between and impact of participants at different disease stages on the developed features should be investigated.

Second, the features used here that were designed to be sensitive to chorea were strongly correlated with the UHDRS-TMS, but no more so than the non-instrumented C3t time scores were shown to be in chapter 2.

Third, the correlation strength between the generated features and chorea appeared to increase as the high-pass filter applied to the signal they were generated from decreased. The best performing feature used a high pass filter of 7.5Hz. This suggests that chorea may be easiest to detect in signals filtered using a bandpass filter with a high-pass filter of 7.5Hz and a low pass filter of 0.3Hz.

Fourth, the heuristic approach used here to select potential filter frequencies produced good results. By calculating the 95^{th}, 90^{th}, and 75^{th} spectral edge frequencies and using these as high pass filters high performing features were found without having to exhaustively search the feature space. When the frequency bandwidth a movement disorder is contained within is not known, this technique could be used to reduce computing time find good (if not necessarily optimal) features.

Fifth, the inclusion/exclusion of the x-axes (the axis most voluntary movement in the C3t is contained along) from generated signal features had little to no impact on feature performance. Similar work should however be mindful of such potential impacts the protocol of action tasks can have when generating features from acceleration/IMU sensors.

Sixth, for the purpose of estimating chorea, there was only a slight improvement seen when using dual task features (i.e., those generated from both the BTT and CTT) as opposed to single task features (i.e., those generated from the BTT or CTT, but not both). Similarly, there was no notable difference in performance between features generated using the BTT and features generated using the CTT. This supports the proposition in chapter 2 that the cognitive load in the CTT is insufficient to cause a decrease in test performance in patients with more severe symptoms. As has been noted however, this needs to be looked at properly in a dedicated study based around detecting differences between performance of the BTT and CTT. Parallel to this the DTT (which should contain an even greater cognitive load) should be investigated.

Chapter 5: Thesis summary & limitations, future directions, and concluding remarks

5.1 Chapter summary

This chapter aims to summarise this thesis and provide the reader with a short, concise view of its chapters and what should come next for the C3t. To start, Thesis summary & limitations provides a short summary of the preceding chapters. The aims of each chapter are highlighted, the key points/results summarised, and the limitations of the underlying studies noted where appropriate. Next, Future work directions discusses what should come next for the C3t. Two distinct direction types are discussed; C3t Technical Development, which discusses what is next for the technical development of the C3t, and C3t Clinical Development which discusses what is needed for the clinical development of the C3t. Finally, there are a few summary points and final comments in Concluding remarks.

5.2 Thesis summary & limitations

5.2.1 Chapter 1: Background, rationale, and work-packages

As more and more clinical trials for HD are undertaken it is vital that effective clinical assessments are available to show the impact of those trials. In chapter 1 it was shown how current assessment strategies lack the necessary sensitivity to detect small changes over time especially in early-stage patients (Reilmann *et al.*, 2011a; Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012; Mestre, Busse, *et al.*, 2018). Detecting change, where change exists, in early HD is particularly important as many potential therapeutics under development aim to slow the progression of HD targeting the disease in its earliest stages.

There is no single optimal method to measure progression in HD. The disease is well known to produce a complex array of symptoms, ranging from movement disorders to cognitive decline to behavioural abnormalities all of which combine to produce functional deficits. The literature suggests that, of the current gold-standard assessments used in HD, the UHDRS-TMS is the most sensitive to early-stage disease progression (Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012). However, it has also been suggested that the UHDRS-TMS is not sensitive enough to detect very small changes in progression during the pre-manifest, prodromal, and early-manifest HD stages (Reilmann *et al.*, 2011a; Tabrizi *et al.*, 2011, 2012, 2013; Meyer *et al.*, 2012; Mestre, Busse, *et al.*, 2018). Composite measures have in recent years been gaining popularity and have been shown to be more sensitive to progression in HD than the traditional single symptom domain assessments used in the UHDRS (Long *et al.*, 2017; Schobel *et al.*, 2017). However, these composite measures use the UHDRS-TMS as their motor assessment input vector. As such, the sensitivity of these composite measures will be limited by the lack of early-stage sensitivity known to exist in the UHDRS-TMS. Additionally, it has been argued that the individual motor assessments which make up the UHDRS-TMS are unlikely to be sufficiently sensitive to assess the progression of individual movement disorders (Reilmann *et al.*, 2011a). Due to the limitations of the UHDRS-TMS, there is an ongoing need for the further development of assessments sensitive to movement disorders in HD.

In PD, where motor symptoms also feature prominently and there are similar issues with the existing gold-standard motor symptoms assessment, instrumented assessments have been widely explored as possible alternatives. A recent review of instrumented assessments in PD shows that they have been highly effective for advancing assessment strategies and providing sensitive measures of progression (Rovini, Maremmani and Cavallo, 2017). Despite their success in PD, instrumented assessments have received less attention in HD, possibly owing to HD being a considerably rarer disease and so receiving less research attention in general. A notable exception however are the Q-Motor series of assessments, although only two have published literature showing relationships to specific HD symptoms (Bechtel *et al.*, 2010; Reilmann *et al.*, 2011a). Regardless, the clear utility of instrumented assessments shown in PD suggests they are worth further exploring for HD.

When developing an instrumented assessment of motor function, a natural first question is whether to target general motor dysfunction of a specific movement disorder. Due to the numerous motor symptoms found in HD developing an assessment sensitive to general HD motor dysfunction necessitates the development of an assessment sensitive to all these symptoms simultaneously. Thus, it arguably makes sense to target motor symptoms individually, and then produce a composite measure of general HD motor dysfunction as/when one is required from the individual assessments.

It should be noted that this is essentially what the UHDRS-TMS currently does – it is a composite score formed from multiple assessments of more specific movement disorders.

The next question is which motor symptoms should be targeted first? A key characteristic of HD is the numerous motor symptoms that can present over a patient's lifespan, all of which deserve the attention of research projects (Roos, 2014). However, as many potential treatments seek to target the earlier stages of HD it makes sense to first focus on developing an assessment specific to early-stage HD motor symptoms. Of the various motor symptoms which may develop during HD, chorea is typically one of the first to develop, and is seen in a large percentage of the HD population (McColgan and Tabrizi, 2018). Additionally, the nature of chorea, which is characterised by sudden involuntary movements is naturally suited to being assessed using simple sensors such as accelerometers and IMUs. Chorea of course also holds a special place in the history of HD, with the disease originally being called Huntington's Chorea (Huntington, 1967; Vale and Cardoso, 2015).

It should be noted that an instrumented assessment designed for chorea already exists, choreomotography, one of the Q-Motor series of assessments. Choreomotography has a participant hold a weighted object stationary for 35 seconds whilst IMUs embedded in the object record velocity. The recorded data allows the total change in position & rotation to be calculated. Reilmann *et al.*, (2011) showed that features produced by the choreomotography assessment are at best a moderately correlation with the chorea section of the UHDRS motor assessment (r=0.48). This moderate correlation between choreomotography and chorea may explain why in a longitudinal study no change was detected in choreomotography scores over 12-, 24-, and 36-month periods, although over this time period chorea would be expected to progress in the studied cohort (Tabrizi *et al.*, 2011, 2012, 2013). Thus, whilst an instrumented assessment of chorea has been developed, it is clear there is room for improvement and so developing such an assessment is the primary focus of this thesis.

The recently developed C3t was chosen as the clinical assessment to be instrumented, due to its known relationship with HD motor function and the manner with which it could be instrumented having been previously demonstrated for a different use-case (Bennasar *et al.*, 2018; Clinch *et al.*, 2018). This first necessitated the further analysis of the C3t, partly to check an adequate relationship with chorea was not already present in the non-instrumented version but also to better understand the C3t as a clinical assessment. In order to support the analysis presented in this thesis an RDCP was conceived to facilitate the large-scale collection of instrumented and non-instrumented C3t data and as well as to support C3t-sensor integration.

At the end of chapter 1 three primary work-packages were defined – an expanded evaluation of the C3t, the development & analysis of a RDCP, and the instrumentation of the C3t with a view to generating measures linked to chorea. Chapters 2, 3, and 4 presented the results of each of these work-packages in turn. Their key findings and the importance of those findings along with relevant study limitations will be discussed in the following sections.

5.2.2 Chapter 2: Expanding analysis on the non-instrumented C3t

In chapter 2 an expanded analysis of the non-instrumented C3t was conducted based around four core aims. First, to assess whether there is a link between the C3t scores and chorea, with a view to creating a 'baseline' against which the instrumented C3t could be judged. Second, to build on previous work by looking at the relationship between the C3t scores and the UHDRS-TMS, CUHDRS, and PIN_{HD}. Third, to assess the C3ts stability over short time periods (baseline to 1-month) and long time periods (baseline to 12-months) relative to the UHDRS-TMS and CUHDRS. Fourth, to simplify the C3t by identifying and removing redundant scores.

Data used was pooled across multiple sites and studies. Two versions were used, the original MBT and the more recent C3t. Of the tasks available in the tests, only the BTT and CTT transfer tasks were used. The TTT (from the MBT) and the DTT (from the C3t) were not used in the study due to the task protocols not overlapping across test versions and as such significantly less data being available for these tasks relative to the BTT and CTT (which are present in both test versions).

No relationship was found between the C3t scores and chorea, however the C3t time scores were strongly correlated with the CUHDRS, UHDRS-TMS and PIN_{HD}. It was also found that the C3t time scores could be used to estimate the CUHDRS and UHDRS-TMS with a high degree of accuracy. Additionally, when applied in different studies or at different clinical sites (within the same study i.e a multi-centre study) or with a different version, there was a negligible impact on the regression model coefficients. Of the 14 C3t scores which were studied only the two C3t time scores were found to be worthy of retention. The remaining 12 scores can be safely discarded due to being invariant, being effectively transformations of the time scores, or having no significant relationship with any of the studied measures. Whilst the C3t time scores do not appear to change over 1-month periods, whether they are truly anchored to changes in the studied clinical measures remains unknown. This is because in the cohort studied neither the C3t time scores nor clinical measures changed over time. As such, the sensitivity of the C3t to changes in the studied clinical measures could not be determined in this study.

Although some results were inconclusive, this chapter significantly enhances our understanding of the non-instrumented C3t as a clinical assessment. It is clear now that the C3t in its non-instrumented state is unlikely to be sufficient for assessing individual motor symptoms, such as chorea. Conceptually this makes sense; the C3t time taken scores will be impacted by general motor dysfunction present in the participant. This will occur for different reasons depending on the symptom; bradykinesia will slow the participant down, chorea & dystonia will make dexterous movements harder, oculomotor dysfunction will make tracking difficult, etc. However, determining which one of these symptoms is having the impact (or their relative contributions to performance degradation) does not appear to be possible. Whilst this has only been shown for chorea, it is likely the case for the other HD motor symptoms.

It is notable, although not a primary finding of this chapter, that the assessment site and test version did not appear to have an impact on the C3t time score's suitability for estimating the CUHDRS or UHDRS-TMS. Whilst only initial evidence, this suggests that the C3t is reliable across multiple study sites and so supports the pooling of data from multiple sites. As HD is a rare disease relative to other conditions such as PD, the pooling of data in this manner can be crucial for obtaining the large sample sizes needed to conduct robust statistical analysis.

The strong relationship and high estimation accuracy of C3t with the CUHDRS opens up the possibility of using it as a quick way to approximate this measure. The component tests of the CUHDRS make it infeasible for a study to regularly collect it in a large-scale study. Regular collection could however provide interesting insights into how the disease progresses over time with a high degree of granularity. The C3t, being significantly simpler, could be used by a patient's primary carer to approximate these more complex measures in the home. It could also be used to approximate them prior to arranging a clinical visit to provide an objective 'snapshot' of a patient's current state. This is especially relevant at time of writing with the global covid-19 pandemic in full swing, making clinical data collection and visits problematic. A strong relationship was also found between the C3t and PIN_{HD}, suggesting it may also be possible to estimate this measure if a larger sample size was available.

Whilst the results from chapter 2 are interesting and encourage further development & refinement of the C3t the work does have three key limitations.

First, only two of the three C3t transfer tasks, the BTT and CTT, were studied during the analysis. The third transfer task, the DTT, was omitted due to small amount of data being available (particularly relative to the amount of data available to the BTT and CTT). The reason for this was the BTT and CTT were both included in the first version of the C3t (the MBT), whilst the DTT was only added during the tests second version, the C3t. As such, data from previous studies was available for the BTT and CTT whilst no such data was available for the DTT, significantly limiting the available sample size. The DTT however should be explored in future work, as is discussed in section 5.3.2.3, as the increased cognitive load of the DTT relative to the BTT and the CTT may impact test performance as was originally hypothesised. Additionally, the performance of the DTT should be examined in order to understand whether it provides any benefit over the BTT and CTT in terms of its relationship with the clinical measures observed during this study. Doing so would inform whether or not the DTT should be retained or omitted from the C3t protocol, a question which cannot currently be answered.

Second, the low sample size available for the longitudinal analysis limits robust conclusions. Although we now know the C3t is strongly corelated to clinical measures in cross-sectional data we still know very little about its behaviour over time. To fully understand the C3t longitudinal data must be properly analysed, and this will need to be covered by future work.

Third, the small sample size for the pre-manifest and prodromal cohort meant that neither regression models nor effect sizes for PIN_{HD} were analysed. As such it is unknown whether the C3t is suitable for estimating PIN_{HD} in the same way it is for estimating the CUHDRS and whether changes in PIN_{HD} over short & long time periods are mirrored by changes in the C3t. Due to the high correlation strength between the C3t scores and PIN_{HD} it is likely high-quality regression models could be built given a larger sample size.

Overall, chapter 2 serves to highlight the clear clinical value of the C3t in its non-instrumented form and suggests ways in which it can be simplified to aid clinical uptake. However, it also shows its limitations regarding the assessment of specific motor symptoms indicating a clear rationale for its instrumentation as covered in chapter 4.

5.2.3 Chapter 3: Developing and evaluating the RDCP

In chapter 3 the design, development, and deployment of the RDCP was detailed and discussed. The rationale for including this chapter was that the development of the RDCP is in itself non-trivial and its eventual deployment was responsible for the strength of the findings presented in chapters 2 and 4. Additionally as is noted in chapter 3, using systems like the RDCP essentially is increasingly common and helps to mitigate many of the pitfalls of traditional data collection methods. As such, showing in this thesis what the construction of such a system looks like and discussing what went well and what did not may be of use to researchers conducting similar work in the future.

The RDCP was ultimately deployed and used to collect data across 3 different projects and 4 different sites. Generally speaking, the system was successful and allowed for the collection of high quality, reliable C3t and accelerometer data, roughly tripling the sample size available for analysis. Unfortunately, some data was either not collected due to protocol problems or was unusable due to technical issues with the accelerometers. This highlighted various issues with the design and development of the RDCP allowing for several recommendations to be made.

First, clinicians should be included in the process when selecting sensors. Many sites experienced issues with the accelerometers which is likely the cause of a lot of the missing or unusable data. Whilst from a technical standpoint the sensors used (GeneActiv tri-axis accelerometers) were fit for purpose, they were ultimately not user friendly. This is a classic problem in engineering where the designers of a system assume that the system will be easy to use for the users because they themselves find it easy to use, but this is not always the case.

Second, additional training and a larger number of information sources was provided to the users. A high number of emails were received from users asking simple questions about the operation of the system. This suggests that either more training may have been required or some degree testing of said training should have been conducted to ensure it was properly understood. Additionally, whilst these questions were covered in the provided manual, alternative information sources (e.g., video tutorials) may have made this information more accessible.

Third, the Waterfall development methodology was used to develop the RDCP however the project fell victim to the methodologies flaw – its inflexibility to changing requirements. Whilst in theory research projects are scoped well ahead of time and thus system specification should not alter, in this instance this was not the case. Ultimately the choice between Waterfall and Agile methodologies (the latter of which is robust to changing requirements at potential cost of development speed) is dependent upon context. In studies where the requirements of software are guaranteed not to change Waterfall is likely to be preferable. However, when requirements may change part the way through the study Agile may be used to reduce the likelihood of unforeseeable increased development costs. In either scenario, the proper scoping of requirements is vital and requires close collaboration by the research and software development teams in order for appropriate software to be delivered in a timely manner.

Fourth, a simple feedback system for the app portion of the RDCP should have been set up. This would have allowed users to report common problems and by analysing this data pain points of the software might have been easier to identify and resolve.

Finally, automatic monitoring of transmitted data should have been implemented. For example, checking accelerometer data was done at time of analysis but should have been completed shortly after the data was received. A common issue was that the C3t and accelerometer data timestamps were misaligned. Similarly, some accelerometer data was corrupted, the recordings containing the correct timestamps but the data itself being unusable. Both of these issues are trivially detectable using appropriate software and should have been automated, with emails being sent out to relevant parties when bad data was detected. This highlights that clinical research will need to adapt typical study protocols as more and more complex technology is integrated. Traditional techniques of managing data quality are not applicable for certain types of data. Similarly, automated data-checking procedures can and should be implemented as more and more data is stored and transmitted in electronic systems and is non-traditional in nature (e.g., acceleration signals).

In general, a more 'clinician-centric' design should have been followed during the RDCPs design and implementation. In particular many problems could have likely been avoided if different sensors had been used. Design cycles are common in software development however the time constraints of this project and the other projects which depended on the RDCP limited the amount that could be practically performed. Overall, along with the recommendations chapter 3 is able to give, it also serves as a reminder that software system construction, which includes hardware that system relies on, truly is non-trivial and should be given proper consideration and time allowance during project planning. Nevertheless, the RDCP was typically successful and was ultimately responsible for the analysis presented in chapters 2 & 4 being possible.

5.2.4 Chapter 4: Assessing the instrumented C3t

In chapter 4 sensor features were generated from accelerometers worn on the wrists of both hands whilst participants took the C3t. In particular two types of features were defined – the number of peaks observed in the signal and the width between these peaks. The rationale for these features was that choreatic movements would be expected to produce a peak in acceleration & jerk signals of the affected limb, and that as chorea worsens choreatic movements occur more regularly & noticeably (Kieburtz *et al.*, 1996). These features were extracted from the instrumented C3t were then used to estimate whole-body chorea, upper-body chorea, and UHDRS-TMS scores. In order to answer several additional questions, numerous subsets of features were produced and analysed as follows.

First, it was felt that jerk, the first derivative of acceleration, might be better suited to producing features sensitive to chorea than the corresponding acceleration signal with respect to the number of peaks in the signal and the widths between them. The rationale for this was rooted in the clinical description of chorea as *"involuntary, unpredictable jerk-like muscle contractions that randomly involve different body parts and vary in frequency, intensity, and amplitude"* (Jankovic and Roos, 2014). Such movements could reasonably be expected to result in an affected limb suddenly accelerating. Jerk, literally defined as the rate of change in acceleration over time, was thus felt to be highly appropriate for assessing the severity of choreatic movements.

Second, as the frequency bandwidth chorea operates within is unknown a heuristic approach was followed to assess the suitability of different bandwidths. Initially four sets of features were generated with different high-pass filters, three of which were based on the cohorts mean 99th, 95th, and 75th

spectral edge frequencies. The best performing features were then re-calculated with different high pass filters ranging from 20Hz to 1Hz with a 0.5Hz step in between them, and the effect of these different filters on the correlation strength of the features with the clinical measures was assessed.

Third, it was theorised that the x-axes, which contains much of the voluntary movement of the C3t (corresponding to horizontal movement across the test board), could obscure the unintentional movements caused by chorea. As such two sets of variants were produced, one set which included the x-axis and one set which did not, and the difference in performance between these sets was assessed.

Fourth, in order to potentially simplify the instrumented C3t by relying on data from only one of the studied transfer tasks three sets of features were developed. The first set of features were generated from both the BTT and CTT (dual task features). The second and third set of features were generated from just the BTT and CTT respectively (single task features). The primary analysis was performed on dual task features, and then the difference in performance between the dual task and single task variants of best performing feature was analysed.

The results from chapter 4 show the benefit that instrumented assessments can have over non-instrumented assessments. Unlike the non-instrumented C3t scores, the generated instrumented C3t features were found to be highly correlated with and predictive of clinical scores of chorea (Spearman's r=0.81; Normalised MAE=15.3%). Notably, this degree of correlation was significantly higher than that reported by the choreomotography Q-Motor assessment (Spearman's r=0.48) (Reilmann *et al.*, 2011a).

The best performing feature (in terms of correlation strength and estimation accuracy with chorea) was the number of peaks in a jerk signal filtered at 7.5Hz. Unlike other studies which rely on combinations of complex features to estimate clinical scores this feature is simple to explain – as chorea increases, the number sudden 'jerky' movements during the instrumented C3t also increases. Although the best performing jerk feature outperformed the best performing acceleration feature the difference was limited. This suggests that whilst there may be some benefit to working with jerk over acceleration for chorea and other hyperkinetic movement disorders it would be wise to generate features from signal types. Additionally, the width between peaks was found to be negatively correlated with increases in chorea. This can be explained that as chorea increases the distance between these jerky movements decreases. The width between peaks was however outperformed by the number of peaks. These findings are however limited due to the lack of video reference data.

Notably none of the instrumented C3t features showed a stronger relationship with the UHDRS-TMS than the C3t time scores. This is likely due to the features being chosen specifically for their hypothesised sensitivity to chorea, whereas the UHDRS-TMS is calculated using multiple movement disorders or which chorea is just one. As such the C3t time scores, which will be affected by any movement disorder that impairs a participant's ability to complete the C3t quickly, would presumably be more sensitive to the UHDRS-TMS than chorea-specific sensor features.

Although they did not outperform the C3t time scores the C3t features were still strongly correlated with the UHDRS-TMS and could estimate it with a reasonable degree of accuracy. This is likely due to the early stage of the studied cohort (during which chorea is prominent) and chorea's heavy weighting in the UHDRS-TMS, rather than the features being sensitive to the multiple movement disorders which comprise the UHDRS-TMS.

Regarding the other types of feature variants, only the filter frequency used had a large impact on feature correlation strength and estimation accuracy. The heuristic approach based on spectral edge frequencies taken in chapter 4 for discovering candidate filter frequencies is computationally inexpensive and may be applied when the optimal filter frequency is unknown. This study serves to highlight the dramatic impact filter frequencies can have on feature performance. Unlike filter frequency the axis and task makeup of the features did not seem to play a large role in feature performance.

A major limitation of this work is the lack of video reference data. Whilst the number of peaks in the jerk signal of the instrumented C3t was positively correlated with chorea, it is unlikely each of these peaks relates to a genuine choreatic movement. Without video data the number of peaks actually caused by choreatic movements cannot be determined. Furthermore, if video data were available the feature could potentially be refined and possibly grant an even greater estimation accuracy of chorea than what is reported here.

Additionally, in a similar manner to the results in chapter 2, only the BTT and CTT were analysed in chapter 4 due to significantly more data being available for these tasks than for the third transfer task, the DTT. Future work should be conducted to assess the usefulness of features generated during the DTT. Parallel to this is the question of whether the CTT, which is thought to cause greater cognitive load than the BTT, impacts developed features in participants with greater degrees of cognitive dysfunction. If the CTT does alter the developed features in such participants relative to their performance in the BTT then this will need to be taken into account in future models and so should be explored in future work.

to this is the question of whether participants with different degrees of cognitive dysfunction are impacted more or less severely in terms of the developed features when taking the CTT which is thought to con

Unfortunately, longitudinal data was not available for the analysis conducted in chapter 4. Whilst one of the studies included in this thesis, PACE-HD, was supposed to include repeated measurements they were not available in sufficient quantity for robust analysis to be conducted. The reason for this was due to the issues some sites experienced with the sensors as discussed in chapter 3. In a similar manner, although the sample size for chapter 4 was reasonable (n=52) ideally many more participants, particularly in the pre-manifest and prodromal stages, would have been available. Increased sample sizes of the pre-manifest and prodromal stages is particularly desirable due to chorea often being one of the first motor symptoms to develop. If an assessment can be developed which can sensitively assess chorea during these stages the tracking of disease progression during said stages could potentially be enhanced.

Finally, this study included participants from a range of disease stages but the impact of doing was not explored as it was felt to be unavoidable and not relevant to the question at hand. However, the relationship between the developed features and HD disease stages is important to consider and so should be explored in future work.

5.3 Future work directions

5.3.1 Section Summary

The chapters presented in this thesis provide a number of interesting avenues for future study. For ease of reading these future work directions are grouped into 2 sections – C3t clinical development and C3t technical development. C3t clinical development details the clinical questions that still need to be investigated to further show the utility of the C3t. C3t technical development deals with how the C3t test, sensors, and RDCP should be developed from a technical standpoint. The high-level ordering of the following sections is not supposed to suggest all clinical development should supersede all technical development. The ordering of the individual items within these sections is however representative of my opinion as to which items should be prioritised. Generally speaking, work items which address fundamental limitations of this thesis (e.g., behaviour of features across disease stages & controls, test-retest effects, trialling different sensors) be prioritised, in my opinion, over more ambitious developments (e.g., looking at additional movement disorders or designing an electronic test board).

5.3.2 C3t Clinical Development

5.3.2.1 Understanding the differences in terms of the developed features across disease stages including controls

The analysis presented in chapter 4 did not include data from control participants, as the goal of the study was to develop features sensitive to chorea rather than distinguish between HD and control participants (as discussed in section 4.3.2.1). Similarly, the work presented in chapter 4 included participants from across the disease stage spectrum but did not explore how the developed features differed across these different stages. As such, future work may wish to investigate how jerk peak & width the features developed in chapter 4 differ across the different HD disease stages and a control group. There are two core reasons for this.

First, it is important to understand the impact (if any) that disease stages are having on the developed features. Doing so can help us improved any models developed in the future (by accounting for the effect on the features) and may inform how the test is continued to be de eloped.

Second, as was discussed throughout chapter 1, the criteria used for staging HD progression in the earliest stages is at best coarse. As the features developed were found to be strongly linked to chorea, commonly one of the first motor symptoms to present, it is possible these features may have use for detecting when motor onset is starting to occur. The first stage of determining whether the features might be useful for this use-case is to determine whether they can be used to distinguish between pre-manifest HD, manifest HD, and control participants.

5.3.2.2 Repeated measures, short-term changes, and long-term changes

A key question about the C3t which remains unanswered by this thesis is how the non-instrumented and instrumented C3t scores change (or do not change) over time. Similarly, it is also unknown whether there is any effect on score performance when the test is taken by the same participant repeatedly. Whilst initial evidence was provided that the C3t is reliable across different sites, this needs to be investigated more fully with a dedicated study.

As has been stated throughout this thesis, understanding how the C3t behaves over time and whether it changes in a similar manner to change in clinical scores is important. In chapter 2 this effect was somewhat studied; however, the results were inconclusive and as such future work should look to see how the C3t scores change over short (e.g., 1-month) and long (e.g., 12-month) time periods. Similarly, understanding whether there is a test-retest effect (i.e., repeated measures) on both the non-instrumented and instrumented C3t scores should be assessed. It is possible participant performance will improve the more a participant takes the test, and this will need to be accounted for if it found to be the case.

Finally, determining whether there is any variation in C3t performance between different sites is also crucial to the C3ts continued development. As conclusively proving the lack of any effect may be difficult (as the sites in question will vary study to study), it is advisable that any future work which pools data from multiple study sites checks for the presence of any such effect during the analysis of any pooled data.

5.3.2.3 DTT exploration and the effect of cognitive load on the developed sensor features

In both chapter 2 and chapter 4 only data from two of the three transfer tasks, the BTT and CTT, were analysed due to data being pooled from multiple studies which included different version of the C3t. The DTT, the final transfer task, is only present in the most recent version of the test, the C3t. As such, significantly less data was available for the DTT relative to the BTT and CTT as these tasks are present in all previous versions of the test. The DTT however contains a significantly greater cognitive load than the BTT and CTT, which is what separates it from the other tasks. It was hypothesised that an increase in cognitive load would elicit greater symptoms and so decrease test performance. However, in this study it was found that there was little difference between the BTT and CTT task times in the context of their relationship with studied clinical measures. It is possible that this is due to the CTT not containing a great enough cognitive load to elicit greater symptoms. Therefore, as the DTT contains a significantly greater cognitive load than the CTT, future work may wish to analyse the task in a similar manner to the analysis presented in chapter 2, in order to understand whether the increased cognitive load has any impact on test performance and so relationship to clinical measures of interest. Similarly, it is also possible that the CTT does contain sufficient cognitive load to exacerbate symptoms, but that the effect is only present in participants with greater degrees of cognitive affectedness rather than the general HD population. As such, future work should also seek to understand whether the developed sensor features appear to be impacted during the CTT in participants with greater degrees of cognitive dysfunction relative to an otherwise comparable control group.

5.3.2.4 Additional movement disorders

In chapter 4 it was shown that the instrumented C3t can produce scores highly correlated with and predictive of the UHDRS measure of chorea. HD however produces a wide variety of disabling movement disorders that sensors are likely suitable for assessing. This thesis provides the first evidence that the C3t may be used to investigate specific movement disorders and future work should seek to explore its relationship with other movement disorders both in HD and other diseases. For example, mean/min/max limb velocity at different stages (e.g., task start, token pickup, token transfer) within a task could provide insights into bradykinesia, absolute measures of rotation could be used to assess dystonia, and joint angle variations could give insights into both dystonia and rigidity.

Additionally, eye movement recordings could be taken from sensors implanted into the test board to assess oculomotor dysfunction during the C3t.

5.3.2.5 Trial the C3t in the home and a fixed-time C3t transfer task

In chapter 1 it was noted that in-home monitoring is a long-term goal of the C3t. It is my opinion that this thesis presents enough information to justify the C3ts utility as a clinical assessment deserving of future research. In my opinion, a fascinating possibility is the ability of the C3t to estimate the CUHDRS composite score. It was shown in chapter 2 that the base C3t time scores can be used, in isolation, to estimate the CUHDRS with a high degree of accuracy. Regularly administering the component assessments needed to compute the CUHDRS is impractical at scale due to the expert clinicians required. However, the C3t is simple to administer and could be reasonably expected to be performed by patients' primary carers or families in the home with minimal training. This would allow a series of estimations of the CUHDRS to be produced and the evolution of these approximations assessed over time, potentially allowing for insights into how HD evolves over time. Such a study would first however need to assess (and handle) any test-retest effects currently present in the C3t, which are currently unknown. The instrumented C3t will also ideally be placed one day into the home in a similar study, but different sensors will need to be trialled as currently they appear too difficult for use in a clinical setting, let alone in the home. If the C3t time scores can similarly be used to estimate PIN_{HD}, which based on the strong correlations observed in chapter 2 seems plausible, then such work could also be conducted for PIN_{HD} potentially giving insights into how early-stage HD evolves over time as well.

In a similar manner to placing the C3t in the home, the features generated in this study should also be trialled for in-home use. Further evidence is needed however to know whether the developed features are truly correlated chorea (the evidence for which could be provided by video reference data) and whether they will work with fixed-time data. As such, we suggest a fixed-time C3t transfer task be produced, with participants being asked to transfer tokens into a container over a fixed-time period. Data collected from this task will more closely mirror real-world in-home monitoring as the amount of collected data will not vary participant-to-participant and could be used to provide evidence that the features developed may be sensitive to chorea in such datasets.

5.3.3 C3t Technical Development

5.3.3.1 Different sensors & automated sensor data monitoring

If an electrical C3t board is not constructed with sensors housed in the tokens the sensors used during this thesis should nevertheless be replaced. As was noted in chapter 3, various study sites encountered significant difficulty operating the sensors. The difficulty sites had operating the sensors had a substantial impact on the amount of data available for analysis. Similar difficulties can be avoided in the future by using sensors more compatible with a clinical setting.

Parallel to the difficulty of actually using the sensors there was also an issue with how the sensor data was monitored. Typically speaking, research data is collected in a database and then manually checked for being present and correct. Acceleration data (and similar digital recordings) are however different to traditional clinical data and require more advanced checking. In this thesis, checking the validity of acceleration data was a two-step process. First, the timestamps of the sensor data and clinical assessment must be checked for synchronicity which includes handling timestamp unit conversions and time-zone differences. Second, the data must be visually assessed to make sure it is not just a flat line, showing significant drift or is unexpectedly noisy. Both of these steps are hard to do manually but could likely be automated.

Future similar work should look to develop a software script that notices when new sensor data has been added to a study folder. Upon noticing a new entry, the database can be queried for corresponding C3t data (e.g., using as participant identifier code). The script should then isolate the correct timestamps in the sensor data for the C3t instance and if it is not found an email can be sent to the study manager. If the correct timestamps are found, then a line plot showing the recorded data can be generated along with various summary metrics and again sent to the study manager for approval or issue flagging.

It is also worth noting that future work make wish to consider the inclusion of IMUs into the instrumented C3t. Whilst accelerometers were found to produce simple, highly correlated features with chorea, IMUs may be needed to develop features sensitive to other motor symptoms seen in HD. For example, dystonia is characterised by repetitive twisting motions and bradykinesia by reduced movement velocity, both of which might be measurable using features derived from IMUs which could not be derived using accelerometers (e.g., joint angles and limb velocity). IMUs also have the advantage of containing accelerometers, and so could still be used to record the features seen to be related to chorea which were developed here. Thus, future work which wishes to enhance the number of motor symptoms the C3t is sensitive to should consider using IMUs in place of accelerometers.

5.3.3.2 C3t app issue tracking & reporting

In a similar vein to the automated checking of acceleration data, user issues should be reportable via the C3t app. Several users had issues with using the app correctly and by allowing for them to report issues directly to the developer the ‘pain points’ of the app could have been dealt with significantly faster. Additionally, it is typically good practice to allow for user feedback and comments. Such functionality should be implemented if the app is developed further, although it would be made needless if the electronic C3t board is developed as this could remove the need for an app entirely depending on its design.

5.3.3.3 Electronic C3t

A simple way to update the C3t test kit is to convert it into an electronic board equipped with sensors suitable for scoring test metrics, giving additional insights into test performance, and simplifying the data storage process.

As was shown in chapter 2, it is likely that only the C3t time scores in the non-instrumented C3t need be collected. Currently, the time taken scores are collected by a clinical assessor using the stopwatch feature provided by the C3t app. Instead, future work could use a series of sensors to detect when a task starts (likely via a physical button) indicating the test had started and when a task stops. Such a device would also allow for more complex measures to be extracted. For example, if the test board can notice when tokens have been picked up, then the time between subsequent token pick-ups (and what are they are picked up in) can be determined. The time between token pick-ups may be related to cognitive dysfunction in tasks which contain a cognitive load to a greater degree than total task time is.

From an engineering standpoint this would be simple to achieve. The board could house a microprocessor which is connected by a circuit to the token indented starting positions. The tokens themselves could have metallic contacts in them which when in their starting positions each complete a circuit to the microprocessor. When a participant removes a token from the board the circuit breaks, cutting power to the relevant microprocessor input and so triggering a 'pick-up' timestamp to be logged. As we would know which input had been disconnected it would also be possible to know which token had been picked up. Noticing a token being placed into the box is equally simple, a simple weight sensor would probably suffice.

Similarly, the tokens themselves could also be modified to contain sensors. In chapter 4, it was shown that accelerometers attached to the wrists of participants were suitable for estimating that participants level of chorea with a high degree of accuracy. Instead of using wearable accelerometers the sensors could be housed in the tokens themselves. If IMUs, which can also measure rotation, were included in the tokens then how the token rotates in space could also be computed. This may give information concerning a participant's general dexterity as well and could even be sensitive to the twisting, writhing movements of dystonia.

Finally, in the home an electronic C3t board could simplify data collection. The microprocessors housed in the boards can connect to the internet, either via WIFI or possible via the 4G/5G network over which test data can be transmitted. Each board can be given a unique ID allowing a backend system to log which board, and so which participant, is sending data. Given the simplicity of the C3t such boards could be manufactured and sent out to participants with their primary carers instructing them how to take the test and all other data processing done automatically.

5.4 Concluding remarks

HD is a devastating disease which profoundly impacts the lives of everyone it touches, made all the worse from the lack of available treatments. Potential treatments are however under active investigation, and it is imperative that effective clinical assessments are available to support them. However, as this thesis has discussed, there is a clear need for the available HD assessments to, in many cases, be updated. This is particularly the case for motor assessments where, as the literature shows, the current gold-standard assessment strategy is inadequate for assessing both general motor function and specific motor symptoms to a sufficiently sensitive level. Fortunately, we live during a time when human movement need not be assessed using vision alone but can also be assessed by augmenting clinical assessments with modern sensor technology – instrumented assessments.

Although instrumented assessments have clear potential for the assessment of movement disorders, as has been shown in PD and explored during this thesis, they are however somewhat of a departure from the clinical norm. As such to aid clinical adoption and ease the transition from traditional clinical assessments to more modernised strategies it is vital that clinicians and engineers work together to develop this new generation of assessments. This collaboration however goes both ways with one of the biggest traps engineers can fall into being to operate in a vacuum and so failing to take into account real clinical needs and patient experiences. This is a lesson I was taught very early on by my academic supervisors and adhering to it throughout this project has served me well. Working with clinicians and collecting data from patients directly was not only one of the most enriching experiences of my doctoral studies but also made the final output, I feel, that much better.

Appendix

6.1 Site and test version coefficient results

Table 45: Lasso regression coefficients for BTT time taken seconds, one-hot encoded sites, and one-hot encoded test versions (C3t & MBT) for estimating UHDRS-TMS and CUHDRS

Regression Feature	UHDRS-TMS		CUHDRS	
	Coefficient Mean	**Coefficient Std**	**Coefficient Mean**	**Coefficient Std**
BTT time taken in seconds	40.45	2.80	-1.51	0.97
Site 1	-5.50	2.28	0.00	0.01
Site 2	0.13	0.40	0.00	0.00
Site 3	0.00	0.00	0.00	0.00
Site 4	-0.71	1.10	0.00	0.00
Site 5	0.00	0.00	0.00	0.00
Site 6	0.00	0.00	0.00	0.00
Site 7	0.40	0.83	0.00	0.00
Site 8	0.00	0.00	0.00	0.00
Site 9	0.00	0.00	0.00	0.00
Site 10	1.24	1.48	0.00	0.00
C3T	0.02	0.19	0.00	0.00
MBT	0.00	0.00	0.00	0.00

Table 46: Lasso regression coefficients for CTT time taken seconds, one-hot encoded sites, and one-hot encoded test versions (C3t & MBT) for estimating UHDRS-TMS and CUHDRS

Regression Feature	UHDRS-TMS		CUHDRS	
	Coefficient Mean	**Coefficient Std**	**Coefficient Mean**	**Coefficient Std**
CTT time taken in seconds	36.06	4.06	-1.62	0.68
Site 1	-6.04	2.40	0.00	0.00
Site 2	0.11	0.46	0.00	0.00
Site 3	-0.04	0.17	0.00	0.00
Site 4	-1.18	1.30	0.00	0.00

Site 5	0.02	0.16	0.00	0.00
Site 6	0.00	0.00	0.00	0.00
Site 7	0.25	0.72	0.00	0.00
Site 8	0.00	0.00	0.00	0.00
Site 9	0.00	0.00	0.00	0.00
Site 10	0.90	1.34	0.00	0.00
C3T	0.00	0.03	0.00	0.00
MBT	0.00	0.00	0.00	0.00

Table 47: Lasso regression coefficients for BTT total task score, one-hot encoded sites, and one-hot encoded test versions (C3t & MBT) for estimating UHDRS-TMS and CUHDRS

Regression Feature	**UHDRS-TMS**		**CUHDRS**	
	Coefficient Mean	**Coefficient Std**	**Coefficient Mean**	**Coefficient Std**
BTT total task score	-42.68	3.90	2.02	0.97
Site 1	-4.99	2.89	0.00	0.00
Site 2	1.74	1.74	0.00	0.00
Site 3	1.21	1.79	0.00	0.00
Site 4	0.00	0.00	0.00	0.00
Site 5	-0.17	0.38	0.00	0.00
Site 6	0.01	0.07	0.00	0.00
Site 7	0.02	0.09	0.00	0.00
Site 8	0.00	0.00	0.00	0.00
Site 9	0.22	0.61	0.00	0.00
Site 10	-0.28	0.60	0.00	0.00
C3T	-0.14	0.49	0.00	0.00
MBT	0.00	0.00	0.00	0.00

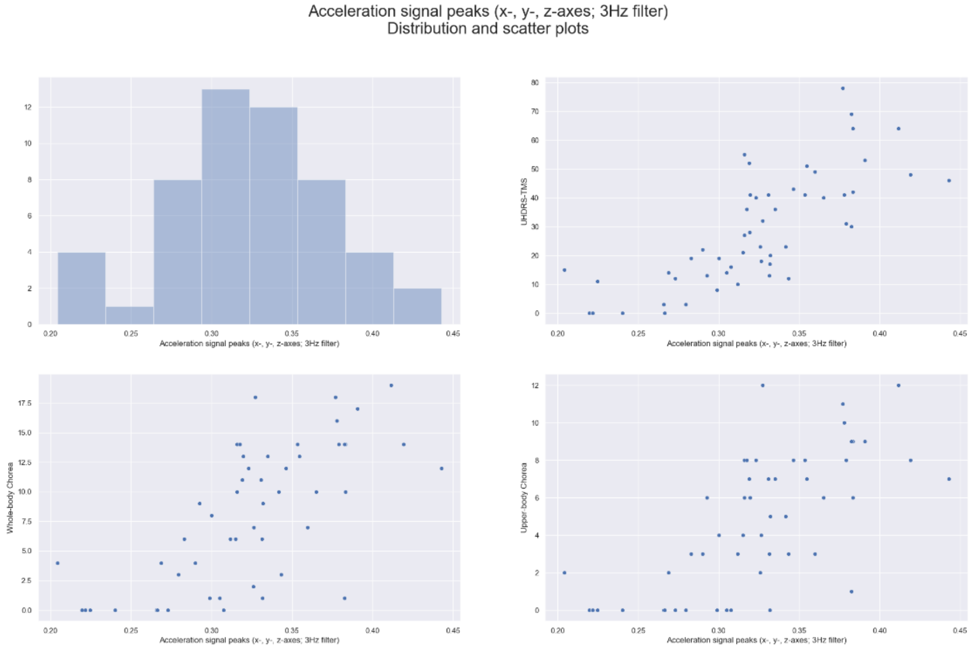
Table 48: Lasso regression coefficients for CTT total task score, one-hot encoded sites, and one-hot encoded test versions (C3t & MBT) for estimating UHDRS-TMS and CUHDRS

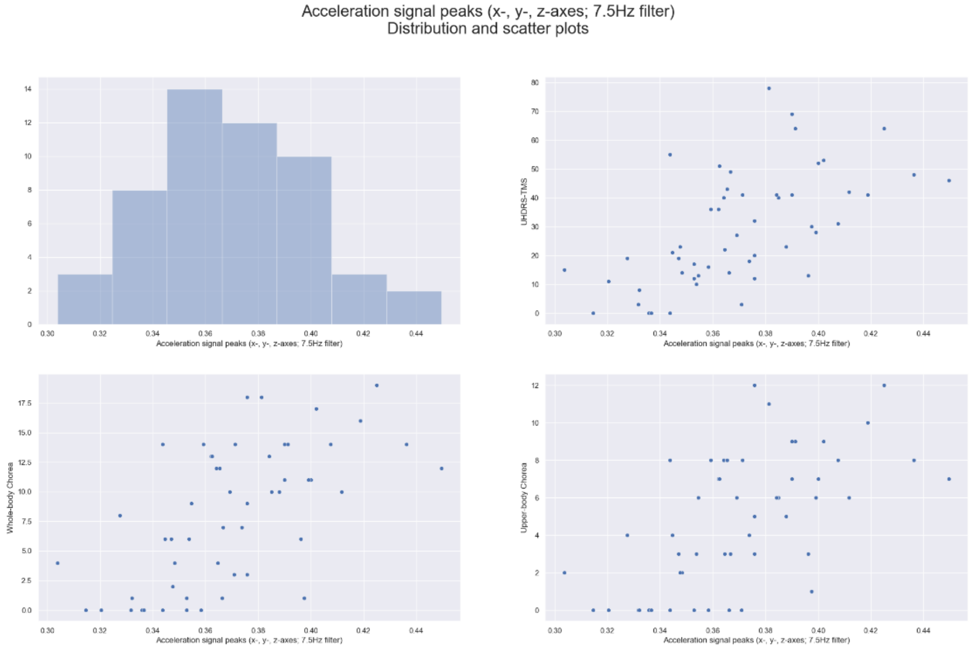
Regression Feature	UHDRS-TMS		CUHDRS	
	Coefficient Mean	**Coefficient Std**	**Coefficient Mean**	**Coefficient Std**
CTT total task score	-41.45	2.52	2.45	0.85
Site 1	-2.27	2.00	0.00	0.00
Site 2	1.35	1.57	0.00	0.00
Site 3	1.67	1.92	0.00	0.00
Site 4	-0.10	0.46	0.00	0.00
Site 5	-0.72	1.04	0.00	0.00
Site 6	0.00	0.00	0.00	0.00
Site 7	-0.03	0.29	0.00	0.00
Site 8	0.00	0.00	0.00	0.00
Site 9	0.00	0.00	0.00	0.00
Site 10	0.00	0.00	0.00	0.00
C3T	-0.02	0.16	0.00	0.00
MBT	0.00	0.00	0.00	0.00

6.2 Chapter 4 Full Results

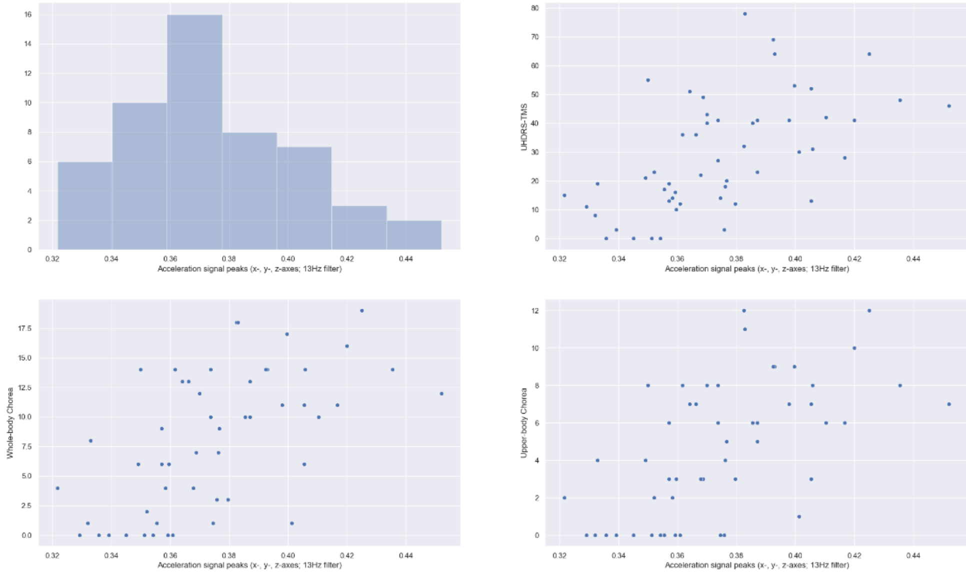
6.2.1 Distribution & scatter plot results

Figure 40: Distribution & scatter plots for each feature. Distribution is shown in the top-left, scatter plot with UHDRS-TMS top-right, scatter plot with whole-body chorea bottom-left, and scatter plot with upper-body chorea bottom right.

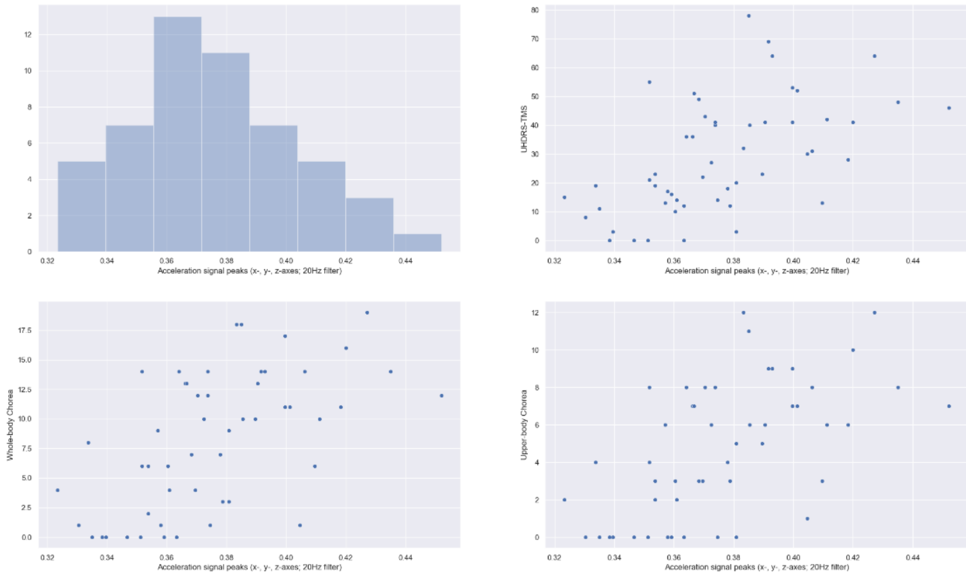




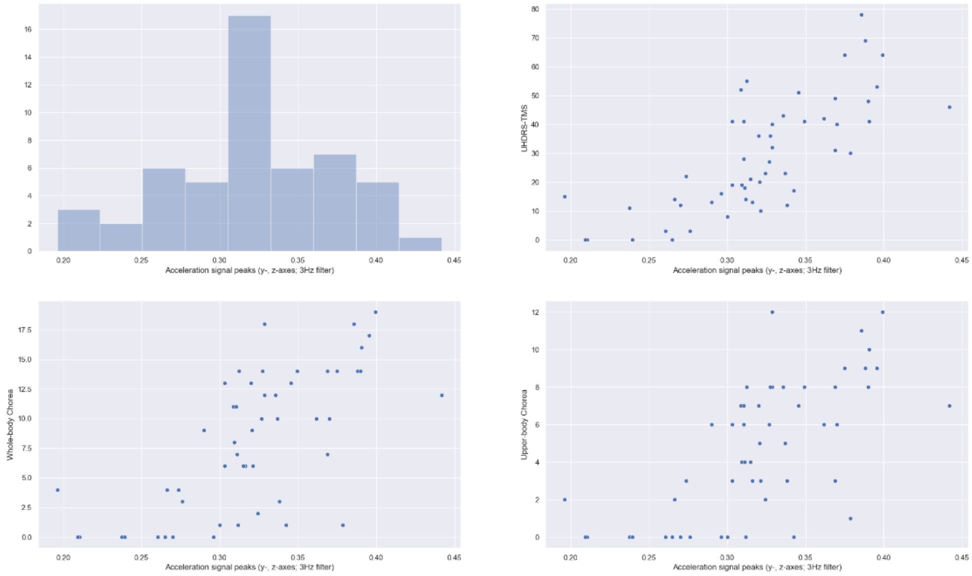
Acceleration signal peaks (x-, y-, z-axes; 13Hz filter)
Distribution and scatter plots



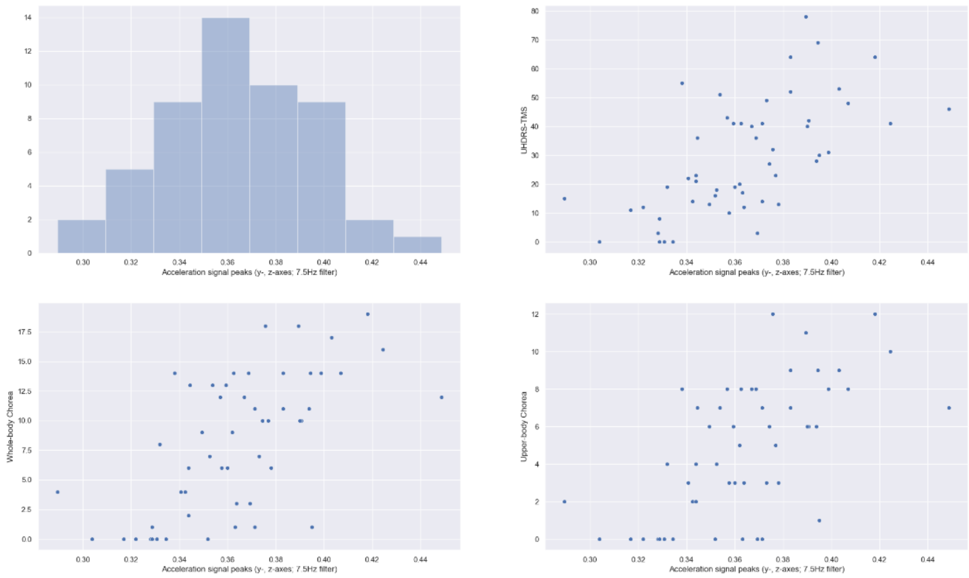
Acceleration signal peaks (x-, y-, z-axes; 20Hz filter)
Distribution and scatter plots



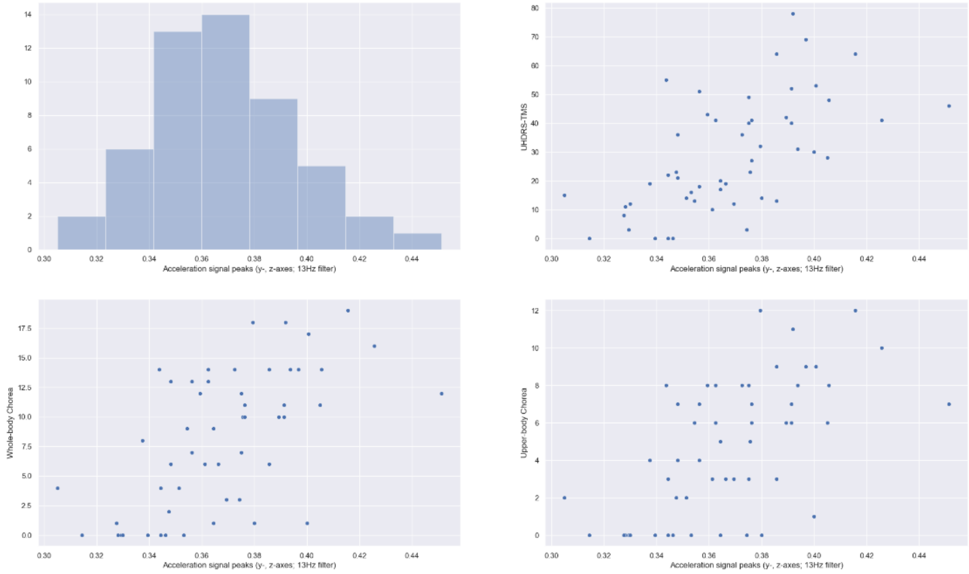
Acceleration signal peaks (y-, z-axes; 3Hz filter)
Distribution and scatter plots



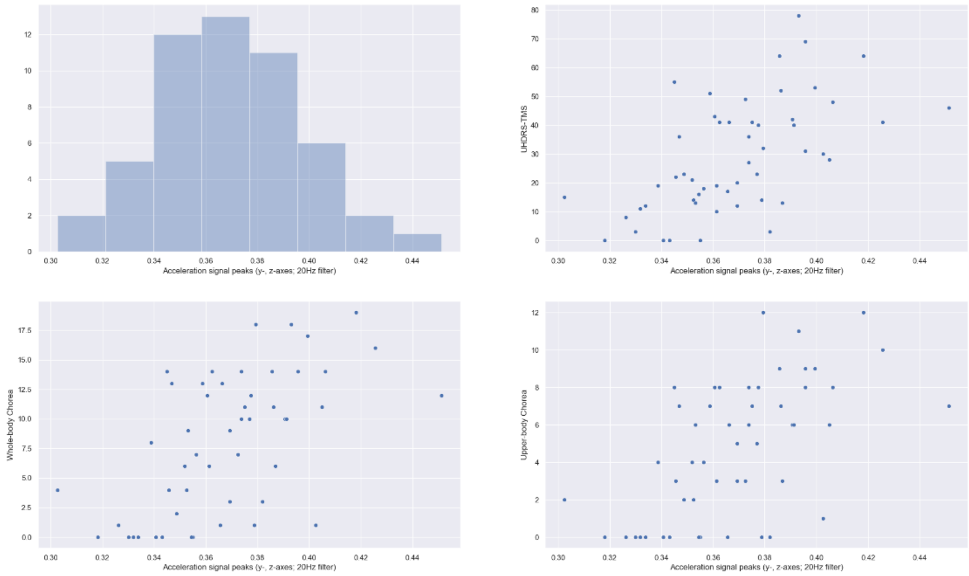
Acceleration signal peaks (y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



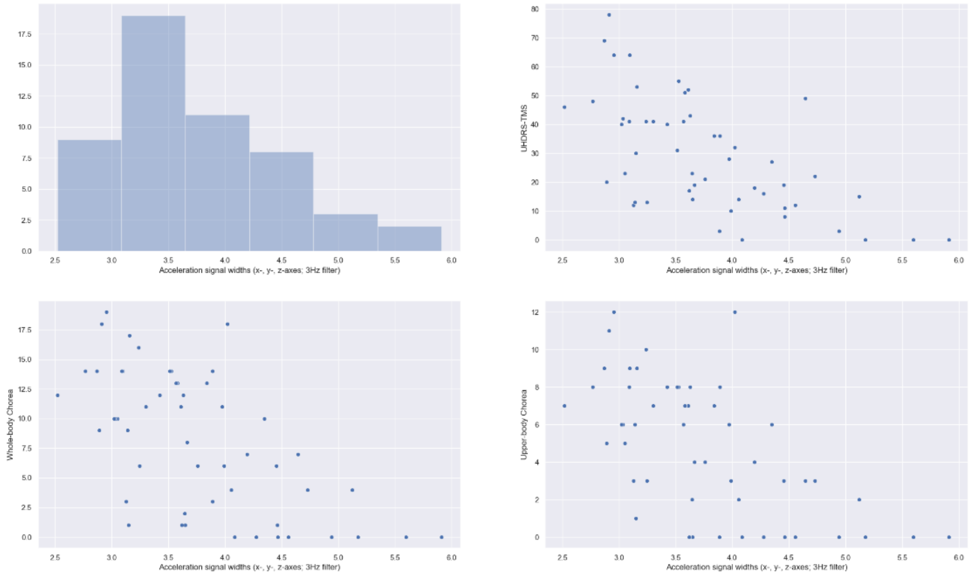
Acceleration signal peaks (y-, z-axes; 13Hz filter)
Distribution and scatter plots



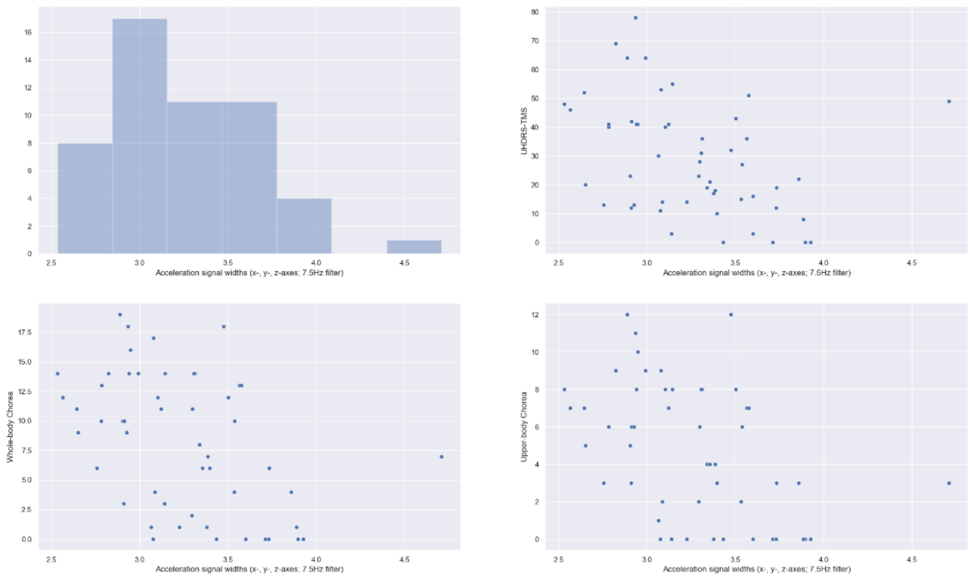
Acceleration signal peaks (y-, z-axes; 20Hz filter)
Distribution and scatter plots



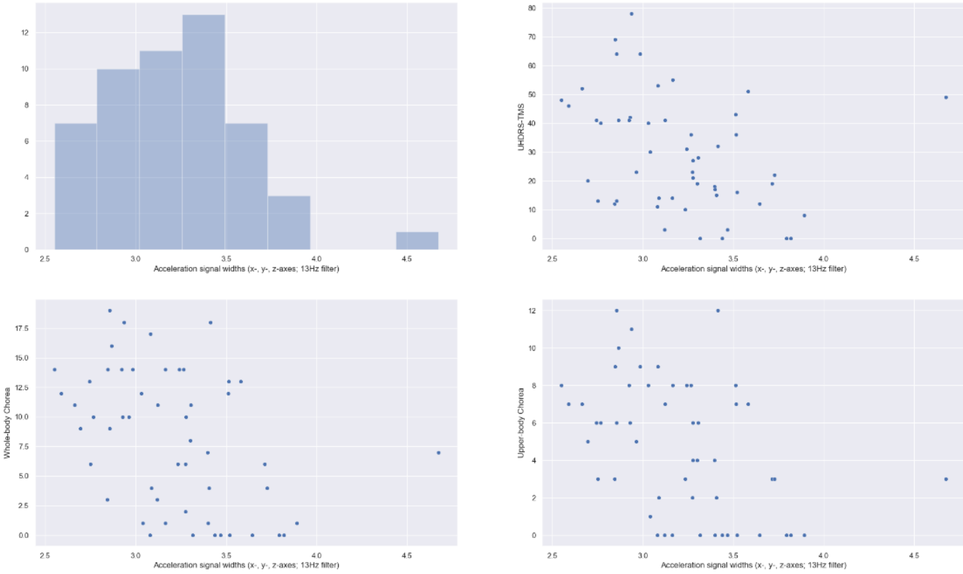
Acceleration signal widths (x-, y-, z-axes; 3Hz filter)
Distribution and scatter plots



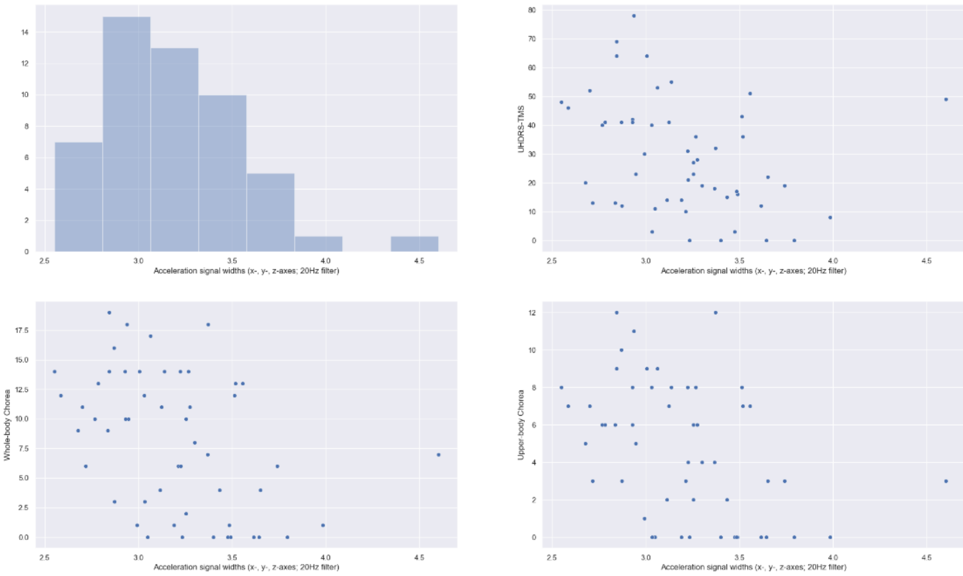
Acceleration signal widths (x-, y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



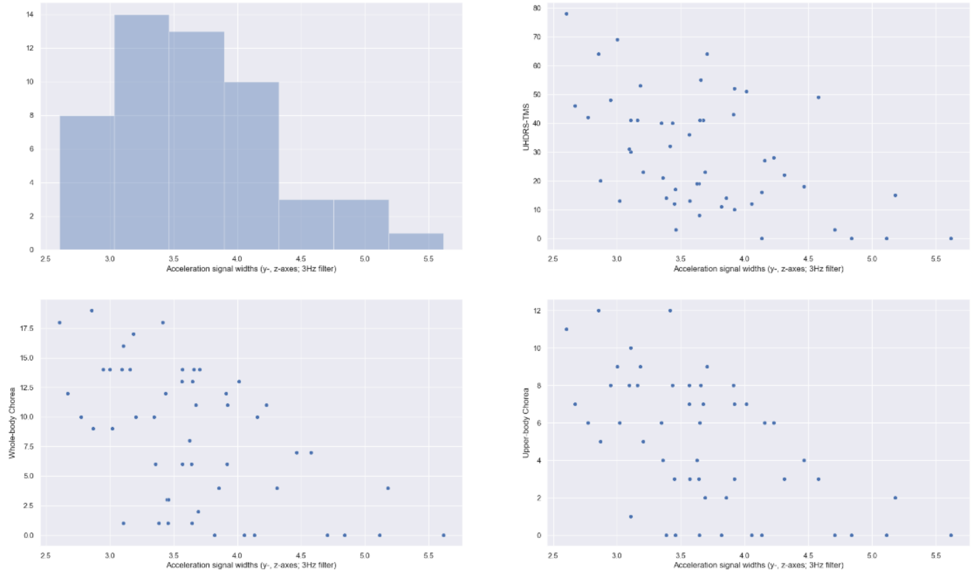
Acceleration signal widths (x-, y-, z-axes; 13Hz filter)
Distribution and scatter plots



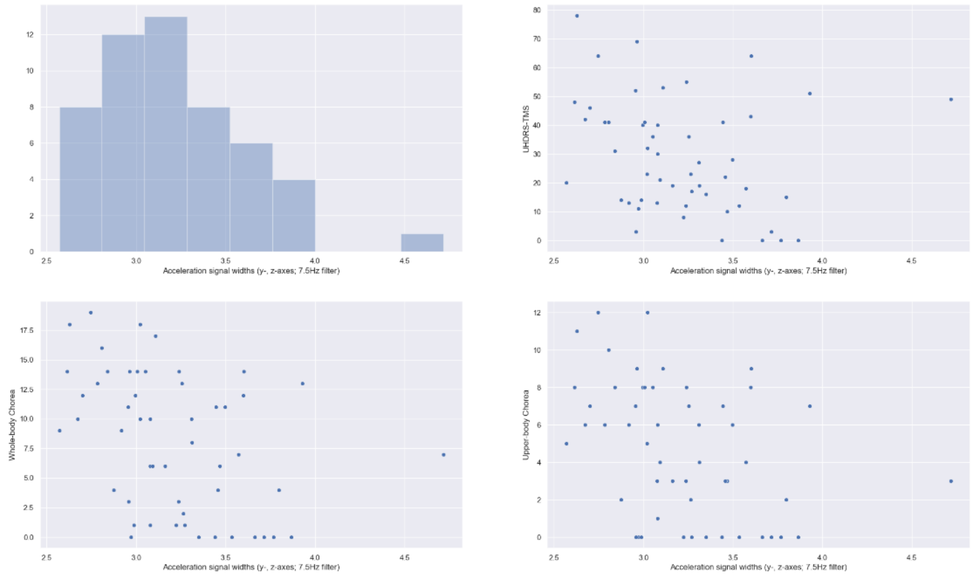
Acceleration signal widths (x-, y-, z-axes; 20Hz filter)
Distribution and scatter plots



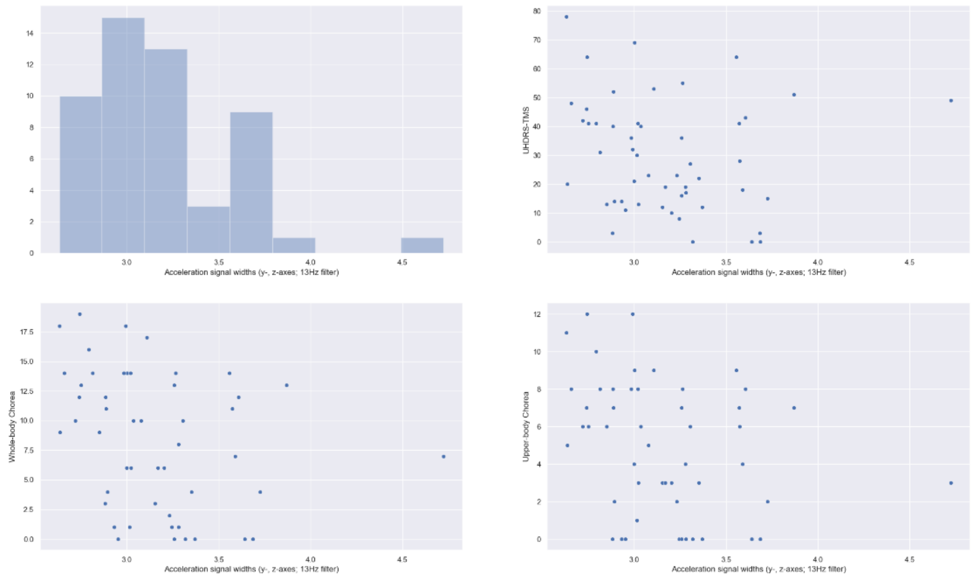
Acceleration signal widths (y-, z-axes; 3Hz filter)
Distribution and scatter plots



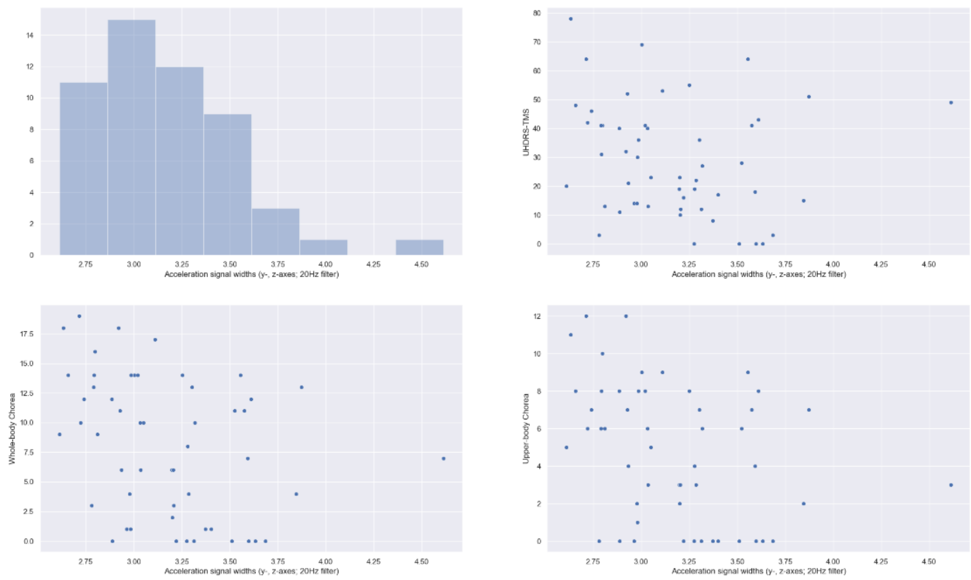
Acceleration signal widths (y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



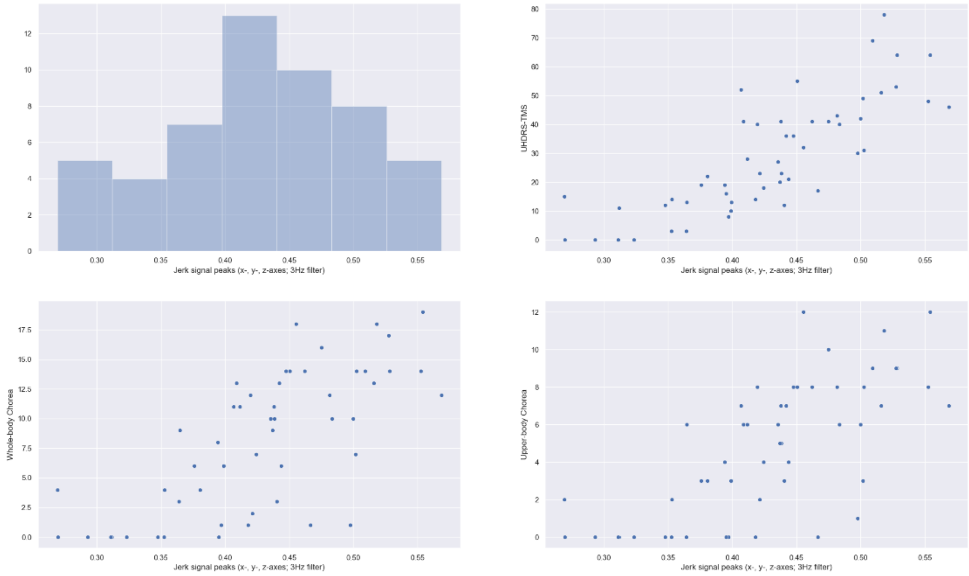
Acceleration signal widths (y-, z-axes; 13Hz filter)
Distribution and scatter plots



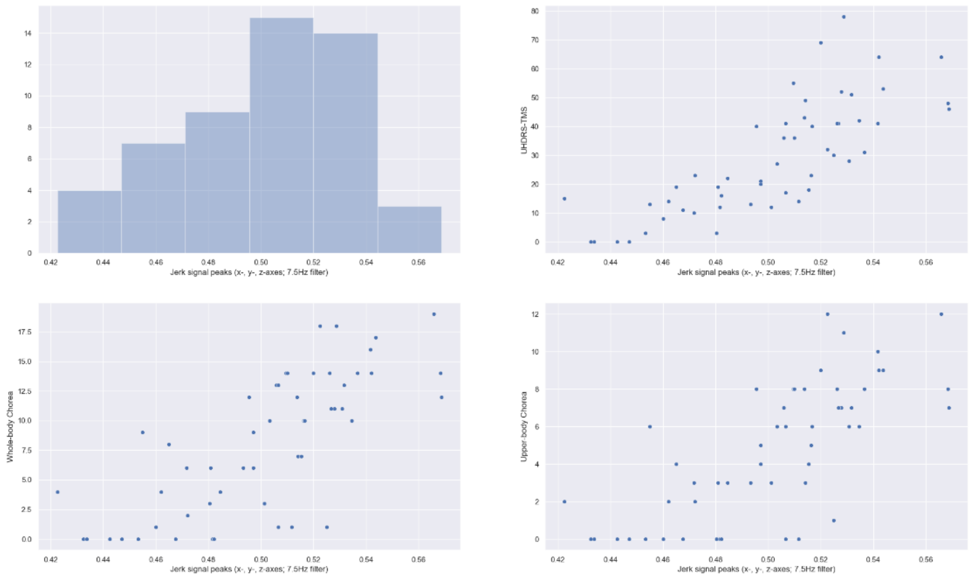
Acceleration signal widths (y-, z-axes; 20Hz filter)
Distribution and scatter plots



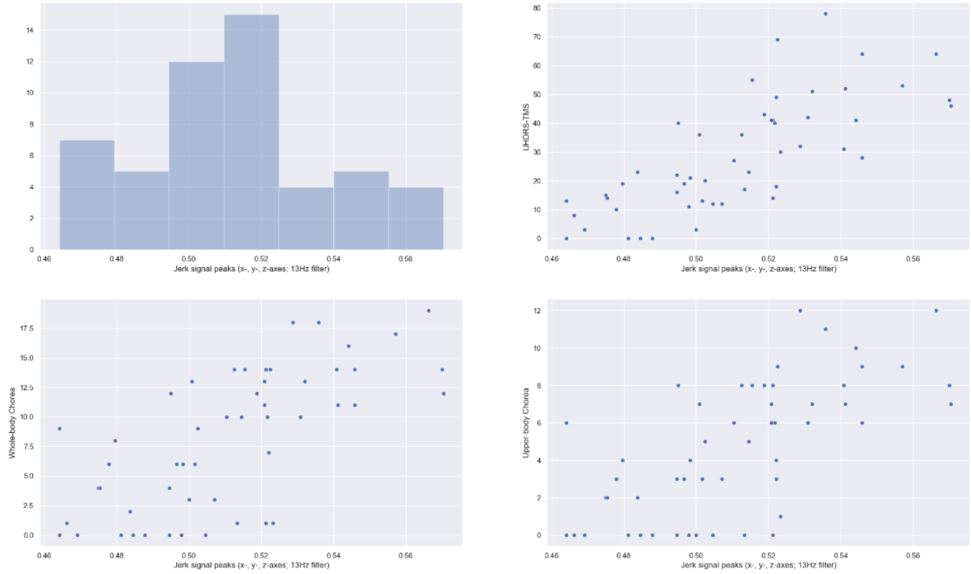
Jerk signal peaks (x-, y-, z-axes; 3Hz filter)
Distribution and scatter plots



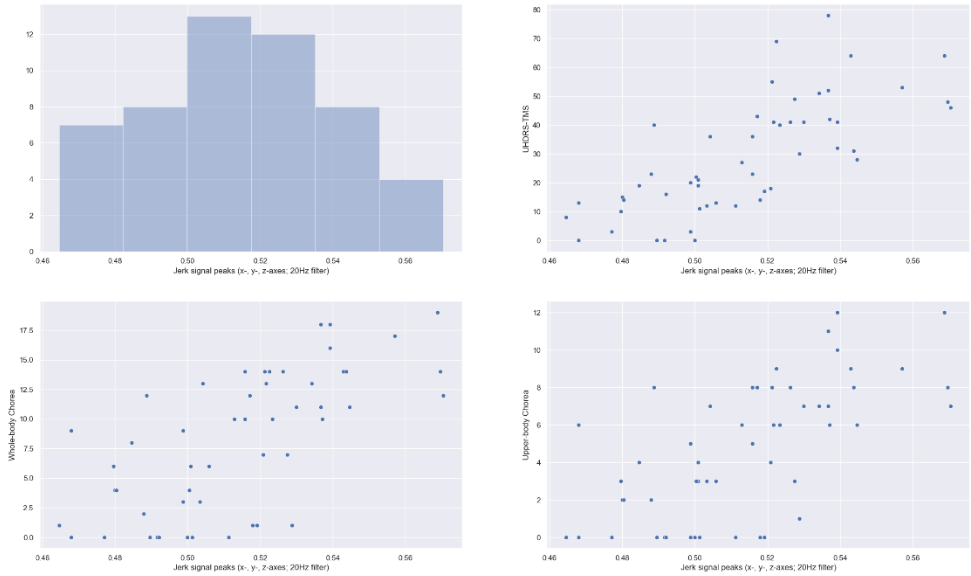
Jerk signal peaks (x-, y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



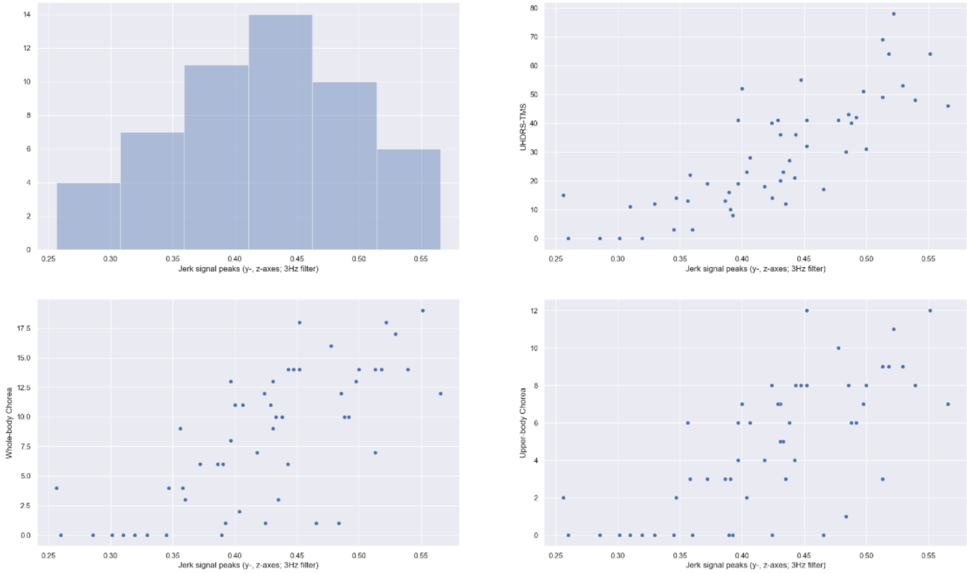
Jerk signal peaks (x-, y-, z-axes; 13Hz filter)
Distribution and scatter plots



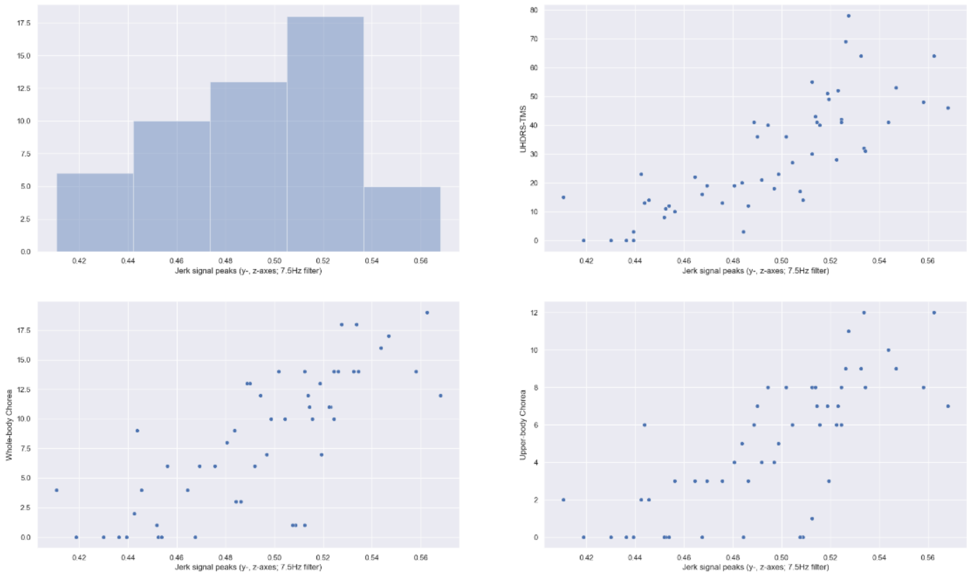
Jerk signal peaks (x-, y-, z-axes; 20Hz filter)
Distribution and scatter plots



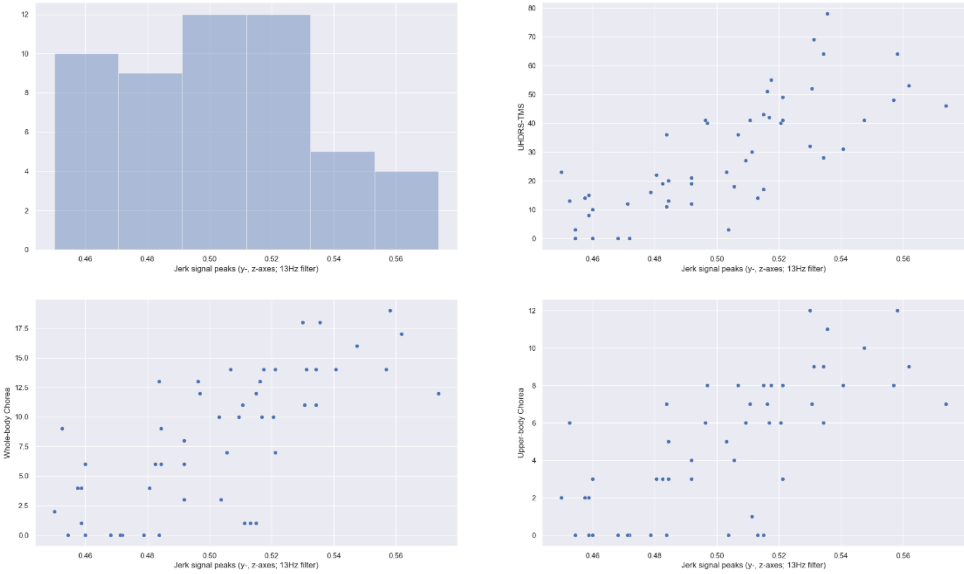
Jerk signal peaks (y-, z-axes; 3Hz filter)
Distribution and scatter plots



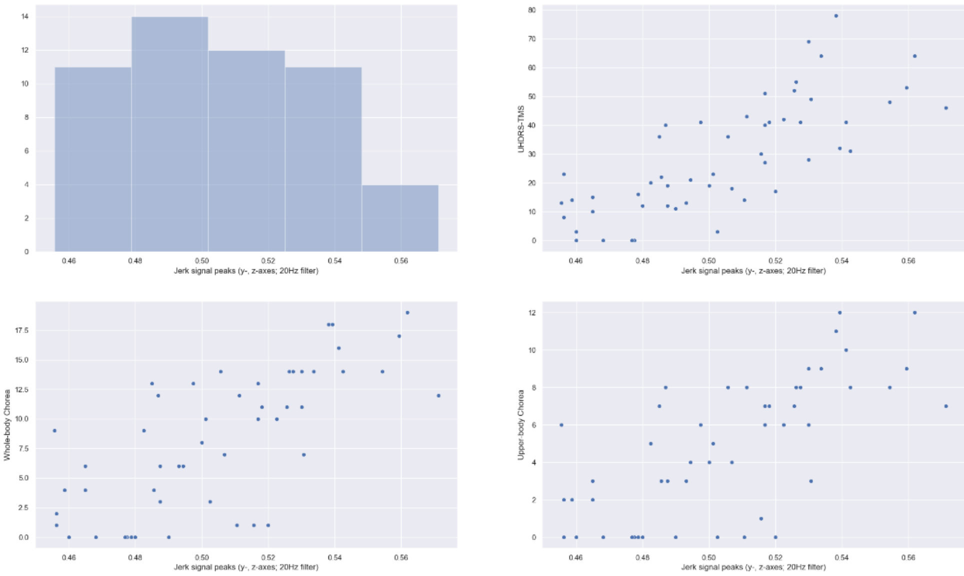
Jerk signal peaks (y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



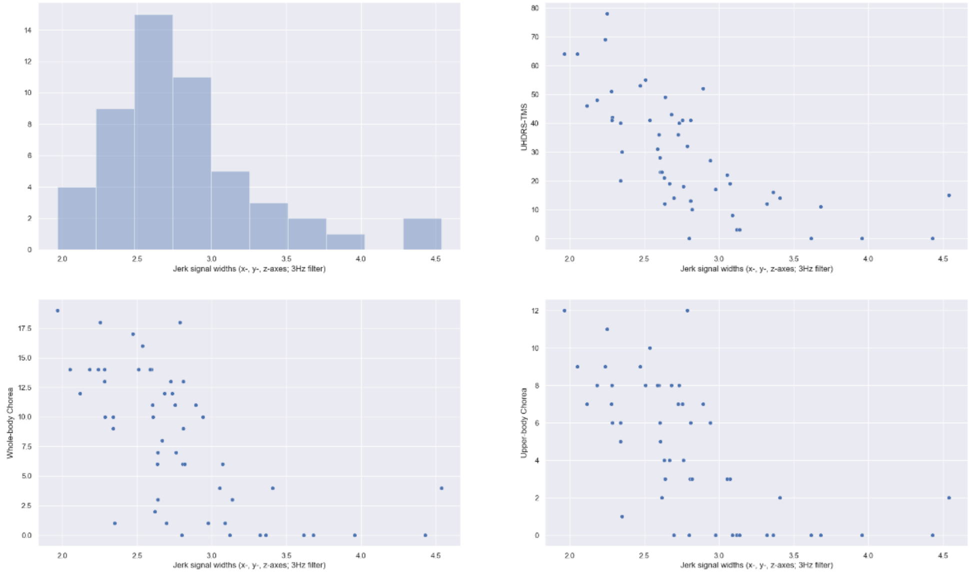
Jerk signal peaks (y-, z-axes; 13Hz filter)
Distribution and scatter plots



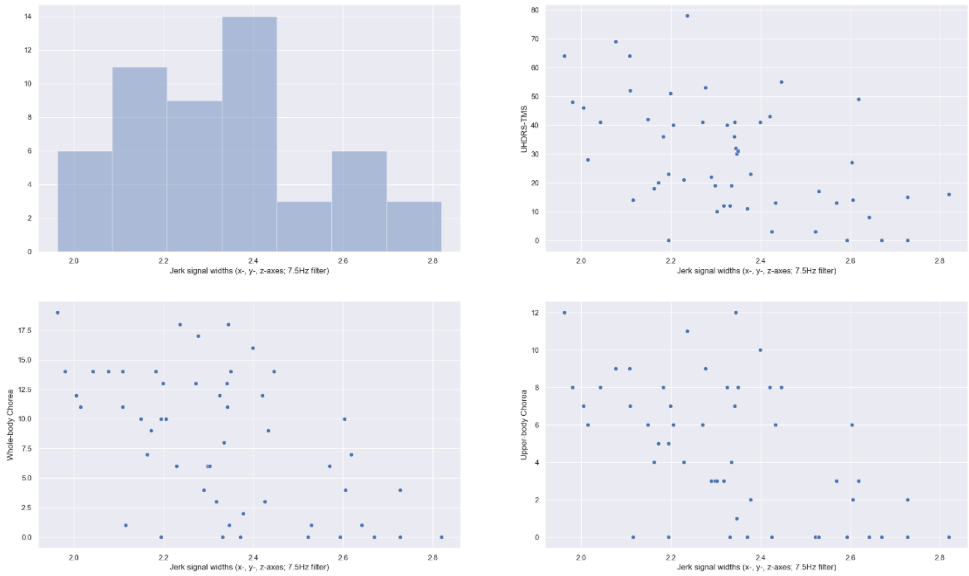
Jerk signal peaks (y-, z-axes; 20Hz filter)
Distribution and scatter plots



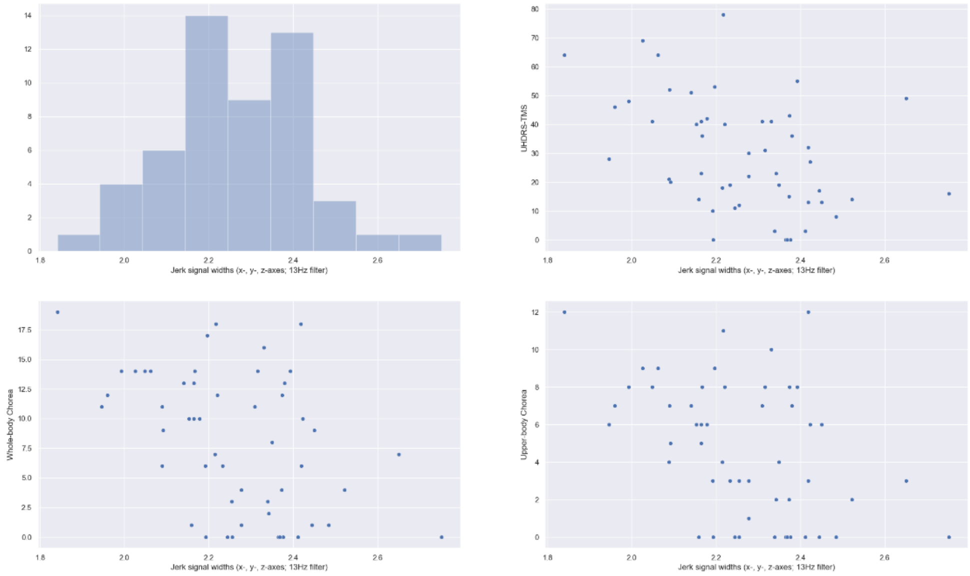
Jerk signal widths (x-, y-, z-axes; 3Hz filter)
Distribution and scatter plots



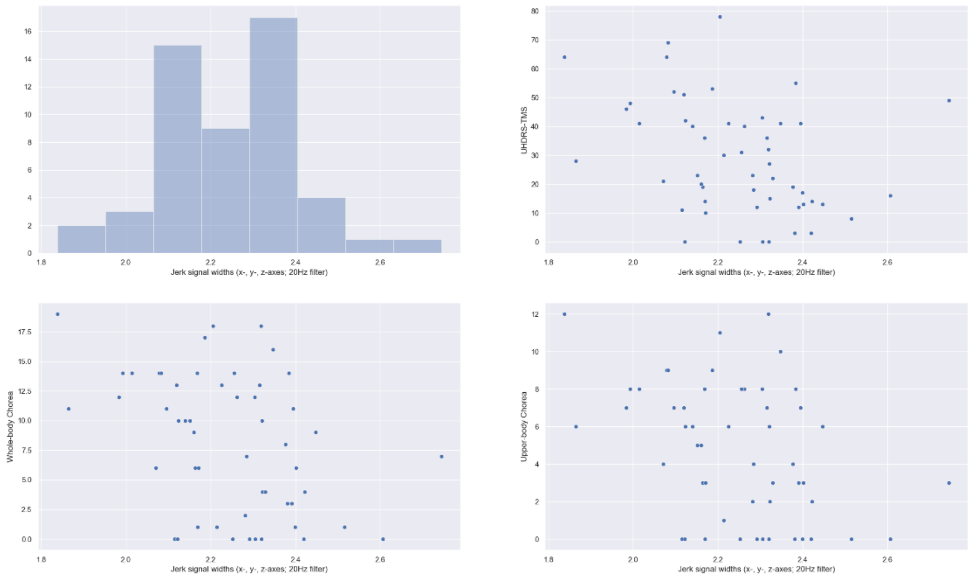
Jerk signal widths (x-, y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



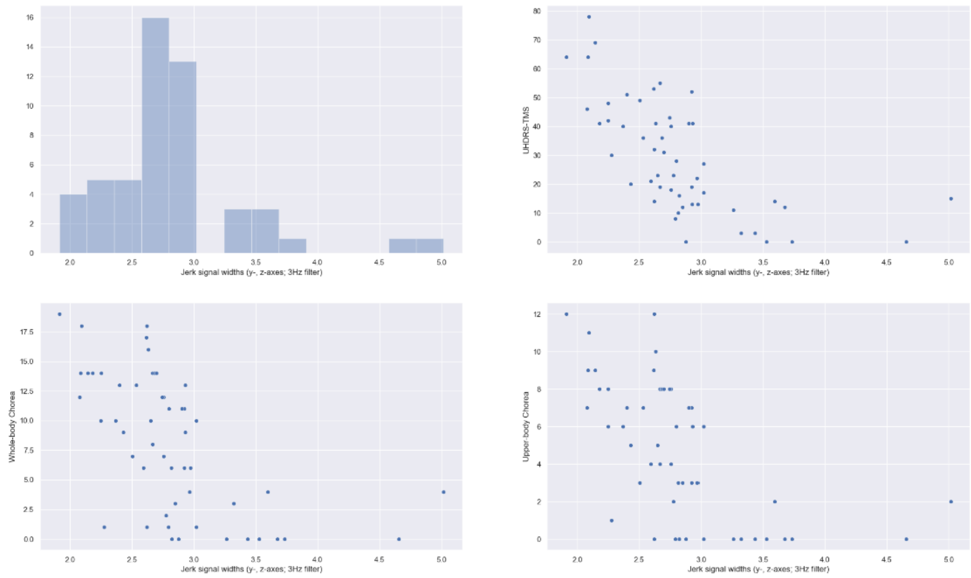
Jerk signal widths (x-, y-, z-axes; 13Hz filter)
Distribution and scatter plots



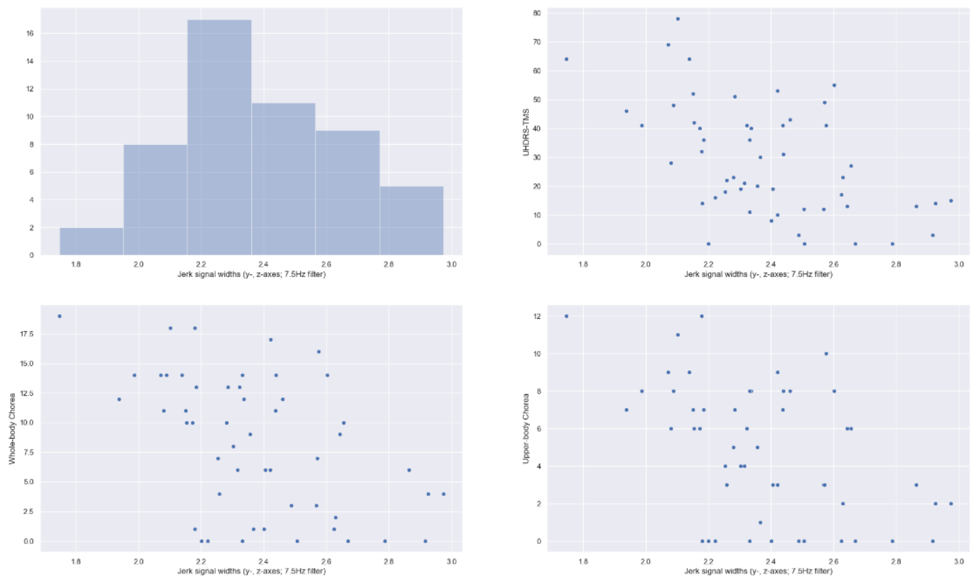
Jerk signal widths (x-, y-, z-axes; 20Hz filter)
Distribution and scatter plots



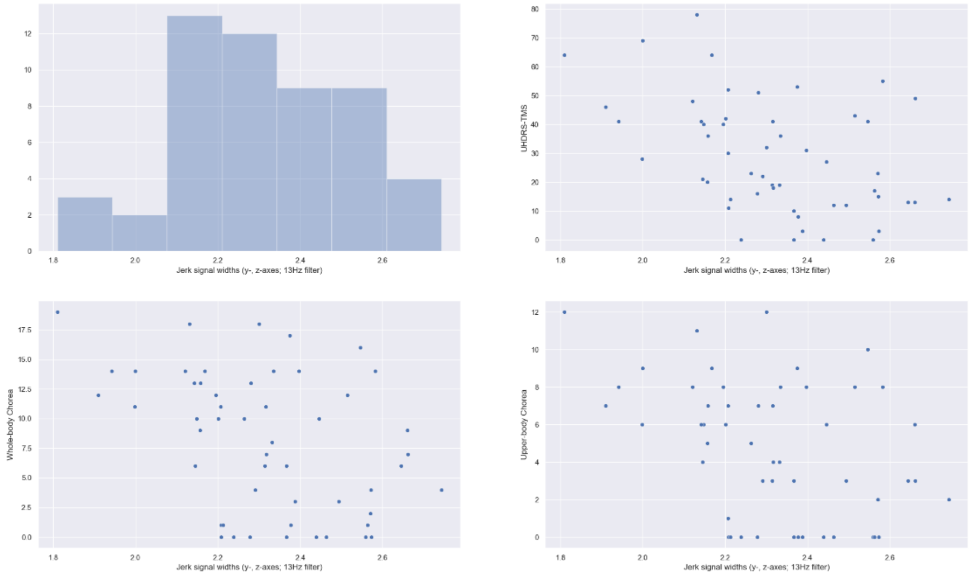
Jerk signal widths (y-, z-axes; 3Hz filter)
Distribution and scatter plots



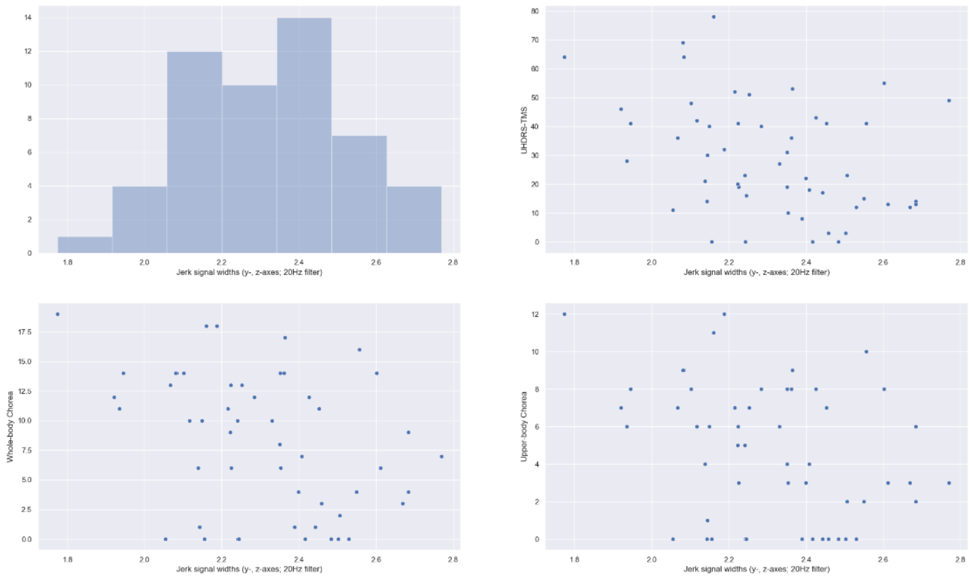
Jerk signal widths (y-, z-axes; 7.5Hz filter)
Distribution and scatter plots



Jerk signal widths (y-, z-axes; 13Hz filter)
Distribution and scatter plots



Jerk signal widths (y-, z-axes; 20Hz filter)
Distribution and scatter plots



6.2.2 Correlation & Regression Results

*Table 49: Correlation & regression results for UHDRS-TMS, *** and ** indicate p<0.001 before Holm-Bonferroni corrections. HB Pass indicates if correlations were significant after Holm-Bonferroni corrections were applied*

UHDRS-TMS				
Feature	**Spearman's R**	**MAE (+-std)**	**Normalised MAE (%)**	**HB Pass?**
Acceleration signal peaks (x-, y-, z-axes; 3Hz filter)	0.79***	10.5 (±1.9)	13.5	pass
Acceleration signal peaks (x-, y-, z-axes; 7.5Hz filter)	0.65***	13.0 (±2.2)	16.7	pass
Acceleration signal peaks (x-, y-, z-axes; 13Hz filter)	0.61***	13.2 (±2.9)	16.9	pass
Acceleration signal peaks (x-, y-, z-axes; 20Hz filter)	0.60***	13.3 (±2.7)	17.1	pass
Acceleration signal peaks (y-, z-axes; 3Hz filter)	0.76***	10.5 (±2.3)	13.5	pass
Acceleration signal peaks (y-, z-axes; 7.5Hz filter)	0.66***	12.8 (±2.3)	16.4	pass
Acceleration signal peaks (y-, z-axes; 13Hz filter)	0.63***	13.0 (±2.2)	16.7	pass
Acceleration signal peaks (y-, z-axes; 20Hz filter)	0.61***	13.2 (±2.3)	16.9	pass
Acceleration signal widths (x-, y-, z-axes; 3Hz filter)	-0.64***	12.3 (±2.8)	15.8	pass
Acceleration signal widths (x-, y-, z-axes; 7.5Hz filter)	-0.45***	15.1 (±2.6)	19.3	pass
Acceleration signal widths (x-, y-, z-axes; 13Hz filter)	-0.40**	15.3 (±2.6)	19.6	fail
Acceleration signal widths (x-, y-, z-axes; 20Hz filter)	-0.38**	15.8 (±2.7)	20.2	fail
Acceleration signal widths (y-, z-axes; 3Hz filter)	-0.46***	14.3 (±2.4)	18.3	pass
Acceleration signal widths (y-, z-axes; 7.5Hz filter)	-0.32**	16.2 (±3.7)	20.8	fail
Acceleration signal widths (y-, z-axes; 13Hz filter)	-0.29**	17.0 (±3.7)	21.9	fail
Acceleration signal widths (y-, z-axes; 20Hz filter)	-0.27	16.9 (±2.9)	21.7	fail
Jerk signal peaks (x-, y-, z-axes; 3Hz filter)	0.84***	9.0 (±1.9)	11.5	pass
Jerk signal peaks (x-, y-, z-axes; 7.5Hz filter)	0.84***	9.4 (±2.0)	12.1	pass
Jerk signal peaks (x-, y-, z-axes; 13Hz filter)	0.79***	10.8 (±2.4)	13.9	pass
Jerk signal peaks (x-, y-, z-axes; 20Hz filter)	0.78***	10.9 (±2.2)	13.9	pass
Jerk signal peaks (y-, z-axes; 3Hz filter)	0.83***	9.0 (±1.7)	11.5	pass

Jerk signal peaks (y-, z-axes; 7.5Hz filter)	0.85***	9.4 (±1.9)	12.0	pass
Jerk signal peaks (y-, z-axes; 13Hz filter)	0.81***	10.7 (±2.7)	13.7	pass
Jerk signal peaks (y-, z-axes; 20Hz filter)	0.79***	10.7 (±2.1)	13.7	pass
Jerk signal widths (x-, y-, z-axes; 3Hz filter)	-0.77***	11.0 (±2.6)	14.1	pass
Jerk signal widths (x-, y-, z-axes; 7.5Hz filter)	-0.51***	14.0 (±3.2)	18.0	pass
Jerk signal widths (x-, y-, z-axes; 13Hz filter)	-0.44**	15.2 (±3.2)	19.4	pass
Jerk signal widths (x-, y-, z-axes; 20Hz filter)	-0.40**	15.5 (±3.1)	19.9	fail
Jerk signal widths (y-, z-axes; 3Hz filter)	-0.74***	11.6 (±1.9)	14.9	pass
Jerk signal widths (y-, z-axes; 7.5Hz filter)	-0.52***	13.9 (±2.6)	17.8	pass
Jerk signal widths (y-, z-axes; 13Hz filter)	-0.45***	14.6 (±2.9)	18.7	pass
Jerk signal widths (y-, z-axes; 20Hz filter)	-0.36**	15.1 (±3.8)	19.3	fail

*Table 50: Correlation & regression results for Whole-body chorea, *** and ** indicate p<0.001 before Holm-Bonferroni corrections. HB Pass refers to whether the correlation was considered significant after Holm-Bonferroni corrections were applied.*

Whole-body Chorea				
Feature	**Spearman's R**	**MAE (±std)**	**Normalised MAE (%)**	**HB Pass?**
Acceleration signal peaks (x-, y-, z-axes; 3Hz filter)	0.71***	3.5 (±0.7)	18.4	pass
Acceleration signal peaks (x-, y-, z-axes; 7.5Hz filter)	0.64***	3.9 (±0.8)	20.5	pass
Acceleration signal peaks (x-, y-, z-axes; 13Hz filter)	0.62***	4.1 (±0.7)	21.5	pass
Acceleration signal peaks (x-, y-, z-axes; 20Hz filter)	0.61***	4.1 (±0.9)	21.6	pass
Acceleration signal peaks (y-, z-axes; 3Hz filter)	0.70***	3.5 (±0.7)	18.7	pass
Acceleration signal peaks (y-, z-axes; 7.5Hz filter)	0.65***	3.8 (±0.8)	19.8	pass
Acceleration signal peaks (y-, z-axes; 13Hz filter)	0.63***	4.0 (±0.6)	21.0	pass
Acceleration signal peaks (y-, z-axes; 20Hz filter)	0.62***	4.0 (±0.8)	20.8	pass
Acceleration signal widths (x-, y-, z-axes; 3Hz filter)	-0.64***	3.8 (±0.7)	20.2	pass
Acceleration signal widths (x-, y-, z-axes; 7.5Hz filter)	-0.47***	4.5 (±0.9)	23.9	pass
Acceleration signal widths (x-, y-, z-axes; 13Hz filter)	-0.45***	4.6 (±0.9)	24.0	pass
Acceleration signal widths (x-, y-, z-axes; 20Hz filter)	-0.43**	4.6 (±0.8)	24.2	fail
Acceleration signal widths (y-, z-axes; 3Hz filter)	-0.55***	4.3 (±0.7)	22.7	pass
Acceleration signal widths (y-, z-axes; 7.5Hz filter)	-0.44**	4.8 (±0.8)	25.2	pass

Acceleration signal widths (y-, z-axes; 13Hz filter)	-0.41**	5.0 (±0.8)	26.4	fail
Acceleration signal widths (y-, z-axes; 20Hz filter)	-0.38**	5.0 (±0.7)	26.4	fail
Jerk signal peaks (x-, y-, z-axes; 3Hz filter)	0.75***	3.3 (±0.8)	17.4	pass
Jerk signal peaks (x-, y-, z-axes; 7.5Hz filter)	0.77***	3.2 (±0.8)	16.6	pass
Jerk signal peaks (x-, y-, z-axes; 13Hz filter)	0.71***	3.6 (±0.6)	18.9	pass
Jerk signal peaks (x-, y-, z-axes; 20Hz filter)	0.70***	3.7 (±0.7)	19.5	pass
Jerk signal peaks (y-, z-axes; 3Hz filter)	0.75***	3.3 (±0.8)	17.2	pass
Jerk signal peaks (y-, z-axes; 7.5Hz filter)	0.81***	2.9 (±0.7)	15.3	pass
Jerk signal peaks (y-, z-axes; 13Hz filter)	0.76***	3.3 (±0.8)	17.4	pass
Jerk signal peaks (y-, z-axes; 20Hz filter)	0.74***	3.3 (±0.7)	17.3	pass
Jerk signal widths (x-, y-, z-axes; 3Hz filter)	-0.72***	3.5 (±0.7)	18.6	pass
Jerk signal widths (x-, y-, z-axes; 7.5Hz filter)	-0.50***	4.1 (±0.8)	21.6	pass
Jerk signal widths (x-, y-, z-axes; 13Hz filter)	-0.41**	4.6 (±0.8)	24.2	fail
Jerk signal widths (x-, y-, z-axes; 20Hz filter)	-0.37**	4.7 (±0.8)	24.5	fail
Jerk signal widths (y-, z-axes; 3Hz filter)	-0.66***	3.8 (±0.8)	20.2	pass
Jerk signal widths (y-, z-axes; 7.5Hz filter)	-0.47***	4.4 (±0.7)	22.9	pass
Jerk signal widths (y-, z-axes; 13Hz filter)	-0.39**	4.6 (±1.0)	24.1	fail
Jerk signal widths (y-, z-axes; 20Hz filter)	-0.32**	4.7 (±1.0)	24.6	fail

*Table 51: Correlation & regression results for upper-body chorea, *** and ** indicate p<0.001 before Holm-Bonferroni corrections. HB Pass refers to whether the correlation was considered significant after Holm-Bonferroni corrections were applied.*

Upper-body Chorea				
Feature	**Spearman's R**	**MAE (+-std)**	**Normalised MAE (%)**	**HB Pass?**
Acceleration signal peaks (x-, y-, z-axes; 3Hz filter)	0.71***	2.1 (±0.5)	17.7	pass
Acceleration signal peaks (x-, y-, z-axes; 7.5Hz filter)	0.63***	2.4 (±0.5)	20.2	pass
Acceleration signal peaks (x-, y-, z-axes; 13Hz filter)	0.61***	2.5 (±0.5)	20.8	pass
Acceleration signal peaks (x-, y-, z-axes; 20Hz filter)	0.60***	2.6 (±0.4)	21.6	pass
Acceleration signal peaks (y-, z-axes; 3Hz filter)	0.70***	2.2 (±0.4)	18.1	pass
Acceleration signal peaks (y-, z-axes; 7.5Hz filter)	0.64***	2.3 (±0.5)	19.2	pass
Acceleration signal peaks (y-, z-axes; 13Hz filter)	0.62***	2.4 (±0.4)	20.1	pass
Acceleration signal peaks (y-, z-axes; 20Hz filter)	0.61***	2.5 (±0.5)	20.5	pass

Acceleration signal widths (x-, y-, z-axes; 3Hz filter)	-0.64***	2.3 (±0.4)	19.3	pass
Acceleration signal widths (x-, y-, z-axes; 7.5Hz filter)	-0.47***	2.7 (±0.5)	22.3	pass
Acceleration signal widths (x-, y-, z-axes; 13Hz filter)	-0.46***	2.7 (±0.4)	22.4	pass
Acceleration signal widths (x-, y-, z-axes; 20Hz filter)	-0.44**	2.8 (±0.6)	23.3	pass
Acceleration signal widths (y-, z-axes; 3Hz filter)	-0.53***	2.6 (±0.5)	21.7	pass
Acceleration signal widths (y-, z-axes; 7.5Hz filter)	-0.41**	2.9 (±0.5)	24.3	fail
Acceleration signal widths (y-, z-axes; 13Hz filter)	-0.39**	3.0 (±0.6)	24.9	fail
Acceleration signal widths (y-, z-axes; 20Hz filter)	-0.36**	3.0 (±0.5)	24.9	fail
Jerk signal peaks (x-, y-, z-axes; 3Hz filter)	0.74***	2.0 (±0.5)	16.8	pass
Jerk signal peaks (x-, y-, z-axes; 7.5Hz filter)	0.75***	1.9 (±0.4)	15.8	pass
Jerk signal peaks (x-, y-, z-axes; 13Hz filter)	0.68***	2.2 (±0.4)	17.9	pass
Jerk signal peaks (x-, y-, z-axes; 20Hz filter)	0.68***	2.2 (±0.5)	18.3	pass
Jerk signal peaks (y-, z-axes; 3Hz filter)	0.74***	2.0 (±0.4)	16.4	pass
Jerk signal peaks (y-, z-axes; 7.5Hz filter)	0.79***	1.8 (±0.4)	14.8	pass
Jerk signal peaks (y-, z-axes; 13Hz filter)	0.74***	2.0 (±0.5)	16.9	pass
Jerk signal peaks (y-, z-axes; 20Hz filter)	0.71***	2.1 (±0.5)	17.4	pass
Jerk signal widths (x-, y-, z-axes; 3Hz filter)	-0.72***	2.2 (±0.4)	18.2	pass
Jerk signal widths (x-, y-, z-axes; 7.5Hz filter)	-0.50***	2.5 (±0.5)	20.7	pass
Jerk signal widths (x-, y-, z-axes; 13Hz filter)	-0.41**	2.8 (±0.5)	22.9	fail
Jerk signal widths (x-, y-, z-axes; 20Hz filter)	-0.38**	2.8 (±0.5)	23.6	fail
Jerk signal widths (y-, z-axes; 3Hz filter)	-0.66***	2.4 (±0.4)	19.9	pass
Jerk signal widths (y-, z-axes; 7.5Hz filter)	-0.46***	2.6 (±0.6)	22.0	pass
Jerk signal widths (y-, z-axes; 13Hz filter)	-0.39**	2.8 (±0.6)	23.0	fail
Jerk signal widths (y-, z-axes; 20Hz filter)	-0.31**	2.9 (±0.6)	24.1	fail

References

Aas, J. *et al.* (2019) 'Let's encrypt: An automated certificate authority to encrypt the entire web', *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 2473–2487. doi: 10.1145/3319535.3363192.

Acosta-Escalante, F. D. *et al.* (2018) 'Meta-classifiers in huntington's disease patients classification, using iPhone's movement sensors placed at the ankles', *IEEE Access*, 6, pp. 30942–30957. doi: 10.1109/ACCESS.2018.2840327.

Adams, H. *et al.* (2014) 'Standardized Assessment', in *Evidence-Based Treatment for Children with Autism*. Elsevier, pp. 501–516. doi: 10.1016/B978-0-12-411603-0.00025-2.

Ali Babar, M. (2014) 'Making Software Architecture and Agile Approaches Work Together', in *Agile Software Architecture*. Elsevier, pp. 1–22. doi: 10.1016/B978-0-12-407772-0.00001-0.

Anand, V. and Rao, C. M. (2016) 'MongoDB and Oracle NoSQL: A technical critique for design decisions', in *1st International Conference on Emerging Trends in Engineering, Technology and Science, ICETETS 2016 - Proceedings*. doi: 10.1109/ICETETS.2016.7602984.

Andrzejewski, K. L. *et al.* (2016) 'Wearable Sensors in Huntington Disease: A Pilot Study', *Journal of Huntington's Disease*, 5(2), pp. 199–206. doi: 10.3233/JHD-160197.

Avidan, A., Weissman, C. and Sprung, C. L. (2005) 'An internet web site as a data collection platform for multicenter research', *Anesthesia and Analgesia*, 100(2), pp. 506–511. doi: 10.1213/01.ANE.0000142124.62227.0F.

Bachoud-Lévi, A., Massart, R. and Rosser, A. (2021) 'Cell therapy in Huntington's disease: Taking stock of past studies to move the field forward', *STEM CELLS*, 39(2), pp. 144–155. doi: 10.1002/stem.3300.

Bakker, J. P. *et al.* (2019) 'A systematic review of feasibility studies promoting the use of mobile technologies in clinical research', *npj Digital Medicine*, 2(1). doi: 10.1038/s41746-019-0125-x.

Bechtel, N. *et al.* (2010) 'Tapping linked to function and structure in premanifest and symptomatic Huntington disease', *Neurology*, 75(24), pp. 2150–2160. doi: 10.1212/WNL.0b013e3182020123.

Beck, K. *et al.* (2001) *Manifesto for Agile Software Development*.

Bellamy, N. (2015) 'Principles of clinical outcome assessment', in *Rheumatology*. Elsevier, pp. 9–19. doi: 10.1016/B978-0-323-09138-1.00002-4.

Bellman, R. E. (1957) 'Dynamic Programming'.

Bennasar, M. *et al.* (2016) 'Huntington's Disease Assessment Using Tri Axis Accelerometers', *Procedia Computer Science*, 96(September), pp. 1193–1201. doi: 10.1016/j.procs.2016.08.163.

Bennasar, M. *et al.* (2018) 'Automated Assessment of Movement Impairment in Huntington's Disease', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(10), pp. 2062–2069. doi: 10.1109/TNSRE.2018.2868170.

Benton, A. L. (1968) 'Differential behavioral effects in frontal lobe disease', *Neuropsychologia*. doi: 10.1016/0028-3932(68)90038-9.

Bouten, C. V. C. C. V. C. *et al.* (1997) 'A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity', *IEEE Transactions on Biomedical Engineering*, 44(3), pp. 136–147. doi: 10.1109/10.554760.

Callaghan, J. *et al.* (2015) 'Reliability and Factor Structure of the Short Problem Behaviors Assessment for Huntington's Disease (PBA-s) in the TRACK-HD and REGISTRY studies', *The Journal of Neuropsychiatry and Clinical Neurosciences*, 27(1), pp. 59–64. doi: 10.1176/appi.neuropsych.13070169.

Carlozzi, N. E. *et al.* (2014) 'Understanding the Outcomes Measures used in Huntington Disease Pharmacological Trials: A Systematic Review.', *Journal of Huntington's disease*, 3(3), pp. 233–52. doi: 10.3233/JHD-140115.

Casula, E. P. *et al.* (2018) 'Response to the letter to the editor by Reilmann et al referring to our article titled "Motor cortex synchronization influences the rhythm of motor performance in premanifest Huntington's disease"', *Movement Disorders*, 33(8), p. 1371. doi: 10.1002/mds.27471.

Chai, T. and Draxler, R. R. (2014) 'Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature', *Geoscientific Model Development*, 7(3), pp. 1247–1250. doi: 10.5194/gmd-7-1247-2014.

Chang, F. *et al.* (2006) 'BigTable: A distributed storage system for structured data', *OSDI 2006 - 7th USENIX Symposium on Operating Systems Design and Implementation*, pp. 205–218.

Clark, R. *et al.* (2019) 'The potential and value of objective eye tracking in the ophthalmology clinic', *Eye (Basingstoke)*, 33(8), pp. 1200–1202. doi: 10.1038/s41433-019-0417-z.

Clarke, C. E. (2007) 'Parkinson's disease', *BMJ*, 335(7617), pp. 441–445. doi: 10.1136/bmj.39289.437454.AD.

Clinch, S. (2017a) *Developing and evaluating behavioural tasks to assess basal ganglia function*. Cardiff University.

Clinch, S. (2017b) *Developing and evaluating behavioural tasks to assess basal ganglia function*. Cardiff University.

Clinch, S. P. *et al.* (2018) 'Rethinking functional outcome measures: The development of a novel upper limb token transfer test to assess basal ganglia dysfunction', *Frontiers in Neuroscience*, 12(MAY). doi: 10.3389/fnins.2018.00366.

Codd, E. F. (1970) 'A relational model of data for large shared data banks', *Communications of the ACM*, 13(6), pp. 377–387. doi: 10.1145/362384.362685.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, *Technometrics*. Routledge. doi: 10.4324/9780203771587.

Craufurd, D. and Snowden, J. (2014) 'Neuropsychiatry and Neuropsychology', in *Huntington's Disease*, pp. 36–65.

Dalton, A. *et al.* (2013) 'Analysis of gait and balance through a single triaxial accelerometer in presymptomatic and symptomatic Huntington's disease', *Gait and Posture*, 37(1), pp. 49–54. doi: 10.1016/j.gaitpost.2012.05.028.

Deo, R. C. (2015) 'Machine learning in medicine', *Circulation*, 132(20), pp. 1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593.

Deuschl, G., Bain, P. and Brin, M. (1998) 'Consensus Statement of the Movement Disorder Society on Tremor', 13, pp. 2–23.

Dinesh, K. *et al.* (2019) 'A Longitudinal Wearable Sensor Study in Huntington's Disease', *Journal of Huntington's Disease*, 9(1), pp. 69–81. doi: 10.3233/jhd-190375.

Eager, D., Pendrill, A. M. and Reistad, N. (2016) 'Beyond velocity and acceleration: Jerk, snap and higher derivatives', *European Journal of Physics*, 37(6). doi: 10.1088/0143-0807/37/6/065008.

Enroll-HD (2020) *Enroll-HD*. Available at: https://enroll-hd.org (Accessed: 23 December 2020).

Estevez-Fraga, C. *et al.* (2021) 'Composite UHDRS Correlates With Progression of Imaging Biomarkers in Huntington's Disease', *Movement Disorders*, pp. 1–7. doi: 10.1002/mds.28489.

Ferrari, A. *et al.* (2016) 'A Mobile Kalman-Filter Based Solution for the Real-Time Estimation of Spatio-Temporal Gait Parameters', *IEEE Transactions on Neural Systems and Rehabilitation*

Engineering, 24(7), pp. 764–773. doi: 10.1109/TNSRE.2015.2457511.

Feys, P. *et al.* (2017) 'The Nine-Hole Peg Test as a manual dexterity performance measure for multiple sclerosis', *Multiple Sclerosis Journal*, 23(5), pp. 711–720. doi: 10.1177/1352458517690824.

Filippeschi, A. *et al.* (2017) 'Survey of motion tracking methods based on inertial sensors: A focus on upper limb human motion', *Sensors (Switzerland)*, 17(6), pp. 1–40. doi: 10.3390/s17061257.

Franklin, J. D., Guidry, A. and Brinkley, J. F. (2011) 'A partnership approach for Electronic Data Capture in small-scale clinical trials', *Journal of Biomedical Informatics*, 44(SUPPL. 1), pp. S103–S108. doi: 10.1016/j.jbi.2011.05.008.

Fritz, N. E. *et al.* (2016) 'Motor-cognitive dual-task deficits in individuals with early-mid stage Huntington disease', *Gait & Posture*, 49, pp. 283–289. doi: 10.1016/j.gaitpost.2016.07.014.

Fusilli, C. *et al.* (2018) 'Biological and clinical manifestations of juvenile Huntington's disease: a retrospective analysis', *The Lancet Neurology*, 17(11), pp. 986–993. doi: 10.1016/S1474-4422(18)30294-1.

Gao, Q. Bin *et al.* (2008) 'EZ-Entry: A clinical data management system', *Computers in Biology and Medicine*, 38(9), pp. 1042–1044. doi: 10.1016/j.compbiomed.2008.07.008.

Gaßner, H. *et al.* (2020) 'Gait variability as digital biomarker of disease severity in Huntington's disease', *Journal of Neurology*, (0123456789). doi: 10.1007/s00415-020-09725-3.

Gazali, Kaur, S. and Singh, I. (2017) 'Artificial intelligence based clinical data management systems: A review', *Informatics in Medicine Unlocked*, 9(September), pp. 219–229. doi: 10.1016/j.imu.2017.09.003.

Gemino, A. and Parker, D. (2009) 'Use case diagrams in support of use case modeling: Deriving understanding from the picture', *Journal of Database Management*. doi: 10.4018/jdm.2009010101.

Gordon, M. F. *et al.* (2019) 'Quantification of Motor Function in Huntington Disease Patients Using Wearable Sensor Devices', *Digital Biomarkers*, 3(3), pp. 103–115. doi: 10.1159/000502136.

Gwin, J. T. *et al.* (2016) 'Wearable Sensors in Huntington Disease: A Pilot Study', *Journal of Huntington's Disease*, 5(2), pp. 199–206. doi: 10.3233/jhd-160197.

Haerder, T. and Reuter, A. (1983) 'Principles of transaction-oriented database recovery', *ACM Computing Surveys (CSUR)*. doi: 10.1145/289.291.

Harris, P. A. *et al.* (2009) 'Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support', *Journal of Biomedical Informatics*, 42(2), pp. 377–381. doi: 10.1016/j.jbi.2008.08.010.

Hasan, H. *et al.* (2017) 'Technologies Assessing Limb Bradykinesia in Parkinson's Disease', *Journal of Parkinson's Disease*, 7(1), pp. 65–77. doi: 10.3233/JPD-160878.

Herrick, R. *et al.* (2016) 'XNAT Central: Open sourcing imaging research data', *NeuroImage*, 124, pp. 1093–1096. doi: 10.1016/j.neuroimage.2015.06.076.

Hersch, S. M. and Rosas, H. D. (2008) 'Neuroprotection for Huntington ' s Disease : Ready , Set , Slow', 5(April), pp. 226–236.

Hiden, H. *et al.* (2013) 'Developing cloud applications using the e-Science Central platform', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983). doi: 10.1098/rsta.2012.0085.

Hohl, P. *et al.* (2018) 'Back to the future: origins and directions of the "Agile Manifesto" – views of the originators', *Journal of Software Engineering Research and Development*, 6(1). doi: 10.1186/s40411-018-0059-z.

Holm, S. (1979) 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics*, 6(2), pp. 65–70. Available at: http://www.jstor.org/stable/4615733.

Huntington, G. (1967) 'On Chorea', *Archives of Neurology*, 17(3), pp. 332–333. doi: 10.1001/archneur.1967.00470270110014.

Inan, O. T. *et al.* (2020) 'Digitizing clinical trials.', *NPJ digital medicine*, 3, p. 101. doi: 10.1038/s41746-020-0302-y.

Initiative, C. T. T. (2018) 'CTTI Recommendations: Advancing the Use of Mobile Technologies for Data Capture & Improved Clinical Trials', (March), pp. 1–32. Available at: https://www.fda.gov/ForIndustry/ImportProgram/ImportBasics/RegulatedProducts/ucm510630.htm%0Ahttps://www.ctti-clinicaltrials.org/sites/www.ctti-clinicaltrials.org/files/mobile-devices-recommendations.pdf%0Ahttps://www.ctti-clinicaltrials.org/projects/mobi.

International Standards Organization (2004) 'ISO 8601:2004(E) Data elements and interchange formats - Information interchange - Representation of dates and times', *Reference number ISO*.

Jankovic, J. and Roos, R. A. C. (2014) 'Chorea associated with Huntington's disease: To treat or not to treat?', *Movement Disorders*, 29(11), pp. 1414–1418. doi: 10.1002/mds.25996.

Le Jeannic, A. *et al.* (2014) 'Comparison of two data collection processes in clinical studies: Electronic and paper case report forms', *BMC Medical Research Methodology*, 14(1). doi: 10.1186/1471-2288-14-7.

Jensen, D. *et al.* (2018) 'F65 Mobile sensor-based gait analysis provides objective motor assessments in huntington's disease', in *Clinical studies*. BMJ Publishing Group Ltd, p. A63.1-A63. doi: 10.1136/jnnp-2018-EHDN.166.

Kegelmeyer, D. A. *et al.* (2017) 'Quantitative biomechanical assessment of trunk control in Huntington's disease reveals more impairment in static than dynamic tasks', *Journal of the Neurological Sciences*, 376, pp. 29–34. doi: 10.1016/j.jns.2017.02.054.

Kieburtz, K. *et al.* (1996) 'Unified Huntington's disease rating scale: Reliability and consistency', *Movement Disorders*, 11(2), pp. 136–142. doi: 10.1002/mds.870110204.

Krabbe, P. F. M. (2017) 'Validity', in *The Measurement of Health and Health Status*. Elsevier, pp. 113–134. doi: 10.1016/B978-0-12-801504-9.00007-6.

Krishnan, S. and Athavale, Y. (2018) 'Trends in biomedical signal feature extraction', *Biomedical Signal Processing and Control*, 43, pp. 41–63. doi: 10.1016/j.bspc.2018.02.008.

Lanza, M. B. *et al.* (2020) 'Intramuscular Fat Influences Neuromuscular Activation of the Gluteus Medius in Older Adults', *Frontiers in Physiology*, 11(December), pp. 1–7. doi: 10.3389/fphys.2020.614415.

Lapinski, M. *et al.* (2019) 'A Wide-Range, Wireless Wearable Inertial Motion Sensing System for Capturing Fast Athletic Biomechanics in Overhead Pitching', *Sensors*, 19(17), p. 3637. doi: 10.3390/s19173637.

Latha, M., Bellary, S. and Krishnankutty, B. (2014) 'Basics of case report form designing in clinical research', *Perspectives in Clinical Research*, 5(4), p. 159. doi: 10.4103/2229-3485.140555.

Li, M. and Rosser, A. E. (2017) 'Pluripotent stem cell-derived neurons for transplantation in Huntington's disease', in *Progress in Brain Research*, pp. 263–281. doi: 10.1016/bs.pbr.2017.02.009.

Long, J. D. *et al.* (2017) 'Validation of a prognostic index for Huntington's disease', *Movement Disorders*, 32(2), pp. 256–263. doi: 10.1002/mds.26838.

Lu, Z. and Su, J. (2010) 'Clinical data management: Current status, challenges, and future directions from industry perspectives', *Open Access Journal of Clinical Trials*, 2, pp. 93–105. doi: 10.2147/oajct.s8172.

Mann, R. K. *et al.* (2012) 'Comparing movement patterns associated with Huntington's chorea and

Parkinson’s dyskinesia’, *Experimental Brain Research*, 218(4), pp. 639–654. doi: 10.1007/s00221-012-3057-0.

Mannini, A. *et al.* (2015) ‘Hidden Markov model-based strategy for gait segmentation using inertial sensors: Application to elderly, hemiparetic patients and Huntington’s disease patients’, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015-Novem, pp. 5179–5182. doi: 10.1109/EMBC.2015.7319558.

Mannini, A. *et al.* (2016) ‘A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington’s disease patients’, *Sensors (Switzerland)*, 16(1). doi: 10.3390/s16010134.

Martin, J. (1983) *Managing the Data-base Environment*. Prentice-Hall. Available at: https://books.google.co.uk/books?id=ymy4AAAAIAAJ.

Martinez-Manzanera, O. *et al.* (2016) ‘A Method for Automatic and Objective Scoring of Bradykinesia Using Orientation Sensors and Classification Algorithms’, *IEEE Transactions on Biomedical Engineering*, 63(5), pp. 1016–1024. doi: 10.1109/TBME.2015.2480242.

Mathiowetz, V. *et al.* (1985) ‘Adult Norms for the Nine Hole Peg Test of Finger Dexterity’, *The Occupational Therapy Journal of Research*, 5(1), pp. 24–38. doi: 10.1177/153944928500500102.

McAllister, B. *et al.* (2021) ‘Timing and Impact of Psychiatric, Cognitive, and Motor Abnormalities in Huntington Disease’, *Neurology*, 96(19), pp. e2395–e2406. doi: 10.1212/WNL.0000000000011893.

McCabe, D. *et al.* (2011) ‘Practice Effects’, in *Encyclopedia of Clinical Neuropsychology*. New York, NY: Springer New York, pp. 1988–1989. doi: 10.1007/978-0-387-79948-3_1139.

McColgan, P. and Tabrizi, S. J. (2018) ‘Huntington’s disease: a clinical review’, *European Journal of Neurology*, 25(1), pp. 24–34. doi: 10.1111/ene.13413.

McConnell, S. (2004) *Code Complete: A Practical Handbook of Software Construction*. 2nd edn. Redmond, WA: Microsoft Press (Best Practices for Developers). Available at: https://www.safaribooksonline.com/library/view/code-complete-second/0735619670/.

McLauchlan, D. (2018) *Objective assessment of the neuropsychiatric symptoms in Huntington’s Disease.* Cardiff University. Available at: http://orca.cf.ac.uk/122117/.

McNally, G. *et al.* (2015) ‘Exploring the validity of the short version of the Problem Behaviours Assessment (PBA-s) for Huntington’s disease: A Rasch analysis’, *Journal of Huntington’s Disease*, 4(4), pp. 347–369. doi: 10.3233/JHD-150164.

Messer-Misak, K., de Bruin, J. S. and Hanke, S. (2020) ‘A Systematic Approach to Quality Requirement Management in Medical Software.’, *Studies in health technology and informatics*, 271, pp. 129–136. doi: 10.3233/SHTI200088.

Mestre, T. A. *et al.* (2016) ‘Rating scales for behavioral symptoms in Huntington’s disease: Critique and recommendations’, *Movement Disorders*, 31(10), pp. 1466–1478. doi: 10.1002/mds.26675.

Mestre, T. A., Busse, M., *et al.* (2018) ‘Rating Scales and Performance-based Measures for Assessment of Functional Ability in Huntington’s Disease: Critique and Recommendations’, *Movement Disorders Clinical Practice*, 5(4), pp. 361–372. doi: 10.1002/mdc3.12617.

Mestre, T. A., Bachoud-Lévi, A.-C. C., *et al.* (2018) ‘Rating scales for cognition in Huntington’s disease: Critique and recommendations’, *Movement Disorders*, 33(2), pp. 187–195. doi: 10.1002/mds.27227.

Mestre, T. A., Forjaz, M. J., *et al.* (2018) ‘Rating Scales for Motor Symptoms and Signs in Huntington’s Disease: Critique and Recommendations’, *Movement Disorders Clinical Practice*, 5(2), pp. 111–117. doi: 10.1002/mdc3.12571.

Meyer, C. *et al.* (2012) ‘Rate of change in early Huntington’s disease: A clinicometric analysis’, *Movement Disorders*, 27(1), pp. 118–124. doi: 10.1002/mds.23847.

Meyer, E. L. *et al.* (2021) ‘Systematic review of available software for multi-arm multi-stage and platform clinical trial design’, *Trials*, 22(1), p. 183. doi: 10.1186/s13063-021-05130-x.

Mills, D. (2006) ‘Simple Network Time Protocol Version 4 for IPv4, IPv6 and OSI’, *RFC4330*.

Myers, R. H. (2004) ‘Huntington’s Disease Genetics’, *NeuroRx*, 1, pp. 255–262. doi: 10.1016/S0140-6736(10)61988-5.

Nahm, M. L., Pieper, C. F. and Cunningham, M. M. (2008) ‘Quantifying data quality for clinical trials using electronic data capture’, *PLoS ONE*, 3(8), pp. 1–8. doi: 10.1371/journal.pone.0003049.

National Institute for Health Research (2014) *Trials Units Information Systems - System Standards*. London. Available at: http://www.ukcrc.org/research-infrastructure/clinical-trials-units/data-and-information-management-systems/%5Cnpapers3://publication/uuid/E8CD92E5-884F-460F-BCA2-ED37B51AE0A7.

Oriented, O., Programming, O. O. and Oo, M. (2001) ‘Object Oriented Programming Concepts 3.1’, *ACM Sigplan Notices*. doi: 10.1145/323648.323751.

Orth, M. *et al.* (2011) ‘Observing Huntington’s Disease: the European Huntington’s Disease Network’s REGISTRY’, *PLoS Currents*, 2(12), p. RRN1184. doi: 10.1371/currents.RRN1184.

Özcan-Top, Ö. and McCaffery, F. (2019) ‘To what extent the medical device software regulations can be achieved with agile software development methods? XP—DSDM—Scrum’, *The Journal of Supercomputing*, 75(8), pp. 5227–5260. doi: 10.1007/s11227-019-02793-x.

Pallmann, P. *et al.* (2018) ‘Adaptive designs in clinical trials: why use them, and how to run and report them’, *BMC Medicine*, 16(1), p. 29. doi: 10.1186/s12916-018-1017-7.

Palmquist, M. S. *et al.* (2013) ‘Parallel Worlds: Agile and Waterfall Differences and Similarities’, *SEI, Carnegie Mellon University*, (October), pp. 1–101.

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. doi: 10.1002/mds.25864.

Park, Y. S. (2016) ‘Correlation Analysis between Dance Experience and Smoothness of Dance Movement by Using Three Jerk-Based Quantitative Methods’, 26(1), pp. 1–9.

Patel, S. *et al.* (2012) ‘A review of wearable sensors and systems with application in rehabilitation’, *Journal of NeuroEngineering and Rehabilitation*, 9(1), p. 21. doi: 10.1186/1743-0003-9-21.

Paulsen, J. S. *et al.* (2010) ‘Challenges Assessing Clinical Endpoints in Early Huntington Disease’, 25(15), pp. 2595–2603. doi: 10.1002/mds.23337.

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. 689–692. doi: 10.1002/ana.410410521.

Perneger, T. V (1998) ‘What’s wrong with Bonferroni adjustments’, *BMJ*, 316(7139), pp. 1236–1238. doi: 10.1136/bmj.316.7139.1236.

Pulliam, C. L. *et al.* (2014) ‘Continuous in-home monitoring of essential tremor’, *Parkinsonism & Related Disorders*, 20(1), pp. 37–40. doi: 10.1016/j.parkreldis.2013.09.009.

Purcell, N. L. *et al.* (2019) ‘The Effects of Dual-Task Cognitive Interference and Environmental Challenges on Balance in Huntington’s Disease’, *Movement Disorders Clinical Practice*, 6(3), pp. 202–212. doi: 10.1002/mdc3.12720.

Rawlins, M. D. *et al.* (2016) ‘The prevalence of huntington’s disease’, *Neuroepidemiology*, 46(2), pp. 144–153. doi: 10.1159/000443738.

Reilmann, R., Bohlen, S., Klopstock, T., Bender, A., Weindl, A., Saemann, P., Auer, D. P., Ringelstein, Erich B., *et al.* (2010) ‘Grasping premanifest Huntington’s disease - shaping new endpoints for new trials’, *Movement Disorders*, 25(16), pp. 2858–2862. doi: 10.1002/mds.23300.

Reilmann, R., Bohlen, S., Klopstock, T., Bender, A., Weindl, A., Saemann, P., Auer, D. P., Ringelstein, E. Bernd, *et al.* (2010) 'Tongue force analysis assesses motor phenotype in premanifest and symptomatic Huntington's disease', *Movement Disorders*, 25(13), pp. 2195–2202. doi: 10.1002/mds.23243.

Reilmann, R. *et al.* (2011a) 'Assessment of involuntary choreatic movements in Huntington's disease-Toward objective and quantitative measures', *Movement Disorders*, 26(12), pp. 2267–2273. doi: 10.1002/mds.23816.

Reilmann, R. *et al.* (2011b) 'Assessment of involuntary choreatic movements in Huntington's disease-Toward objective and quantitative measures', *Movement Disorders*, 26(12), pp. 2267–2273. doi: 10.1002/mds.23816.

Reilmann, R. and Schubert, R. (2017a) 'Motor outcome measures in Huntington disease clinical trials', in, pp. 209–225. doi: 10.1016/B978-0-12-801893-4.00018-3.

Reilmann, R. and Schubert, R. (2017b) *Motor outcome measures in Huntington disease clinical trials*. 1st edn, *Handbook of Clinical Neurology*. 1st edn. Elsevier B.V. doi: 10.1016/B978-0-12-801893-4.00018-3.

Rescorla, E. (2000) *RFC2818: HTTP Over TLS*. doi: 10.17487/rfc2818.

Romano, P. *et al.* (2007) 'Biowep: A workflow enactment portal for bioinformatics applications', *BMC Bioinformatics*, 8(SUPPL. 1), pp. 1–13. doi: 10.1186/1471-2105-8-S1-S19.

Roos, R. A. C. (2014) *Clinical Neurology*. Oxford University Press. doi: 10.1093/med/9780199929146.003.0002.

Rosa, C. *et al.* (2021) 'Using digital technologies in clinical trials: Current and future applications', *Contemporary Clinical Trials*, 100, p. 106219. doi: 10.1016/j.cct.2020.106219.

Rosser, A. and Svendsen, C. N. (2014) 'Stem cells for cell replacement therapy: A therapeutic strategy for HD?', *Movement Disorders*, 29(11), pp. 1446–1454. doi: 10.1002/mds.26026.

Rovini, E., Maremmani, C. and Cavallo, F. (2017) 'How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review.', *Frontiers in neuroscience*, 11(OCT), p. 555. doi: 10.3389/fnins.2017.00555.

Ruonala, V. *et al.* (2014) 'EMG signal morphology and kinematic parameters in essential tremor and Parkinson's disease patients', *Journal of Electromyography and Kinesiology*, 24(2), pp. 300–306. doi: 10.1016/j.jelekin.2013.12.007.

Russell, C. *et al.* (2021) 'Choosing a Mobile Sensor Technology for a Clinical Trial: Statistical Considerations, Developments and Learnings', *Therapeutic Innovation & Regulatory Science*, 55(1), pp. 38–47. doi: 10.1007/s43441-020-00188-2.

Sánchez-De-Madariaga, R. *et al.* (2017) 'Examining database persistence of ISO/EN 13606 standardized electronic health record extracts: Relational vs. NoSQL approaches', *BMC Medical Informatics and Decision Making*, 17(1), pp. 1–14. doi: 10.1186/s12911-017-0515-4.

Scarpina, F. and Tagini, S. (2017) 'The stroop color and word test', *Frontiers in Psychology*, 8(APR), pp. 1–8. doi: 10.3389/fpsyg.2017.00557.

Schobel, S. A. *et al.* (2017) 'Motor, cognitive, and functional declines contribute to a single progressive factor in early HD', *Neurology*, 89(24), pp. 2495–2502. doi: 10.1212/WNL.0000000000004743.

Schober, P. and Schwarte, L. A. (2018) 'Correlation coefficients: Appropriate use and interpretation', *Anesthesia and Analgesia*, 126(5), pp. 1763–1768. doi: 10.1213/ANE.0000000000002864.

Schulte, J. and Littleton, J. T. (2011) 'The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology.', *Current trends in neurology*, 5, pp. 65–78. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22180703.

Sergios Theodoridis and Koutroumbas, K. (2009) *Pattern Recognition (Fourth Edition)*, *Pattern Recognition*.

Shoulson, I. and Fahn, S. (1979) 'Huntington disease: clinical care and evaluation.', *Neurology*, 29(1), pp. 1–3. doi: 10.1212/wnl.29.1.1.

Smith, A. (1968) 'The Symbol-Digit Modalities Test: A neuropsychologic test for economic screening of learning and other cerebral disorders', *Learning Disorders*.

Solari, A. *et al.* (2005) 'The multiple sclerosis functional composite: different practice effects in the three test components', *Journal of the Neurological Sciences*, 228(1), pp. 71–74. doi: 10.1016/j.jns.2004.09.033.

Sorzano, C. O. S., Vargas, J. and Montano, A. P. (2014) 'A survey of dimensionality reduction techniques', pp. 1–35. Available at: http://arxiv.org/abs/1403.2877.

Steinhubl, S. R. *et al.* (2019) 'Digital clinical trials: creating a vision for the future', *npj Digital Medicine*, 2(1), p. 126. doi: 10.1038/s41746-019-0203-0.

Stroop, J. R. (1935) 'Studies of interference in serial verbal reactions.', *Journal of Experimental*

Psychology, 18(6), pp. 643–662. doi: 10.1037/h0054651.

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', *The Lancet Neurology*, 10(1), pp. 31–42. doi: 10.1016/S1474-4422(10)70276-3.

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. doi: 10.1016/S1474-4422(11)70263-0.

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', *The Lancet Neurology*, 12(7), pp. 637–649. doi: 10.1016/S1474-4422(13)70088-7.

Teshuva, I. *et al.* (2019) 'Using wearables to assess bradykinesia and rigidity in patients with Parkinson's disease: a focused, narrative review of the literature', *Journal of Neural Transmission*, 126(6), pp. 699–710. doi: 10.1007/s00702-019-02017-9.

Thwin, S. S. *et al.* (2007) 'Automated inter-rater reliability assessment and electronic data collection in a multi-center breast cancer study', *BMC Medical Research Methodology*, 7, pp. 1–8. doi: 10.1186/1471-2288-7-23.

Trigili, E. *et al.* (2019) 'Detection of movement onset using EMG signals for upper-limb exoskeletons in reaching tasks', *Journal of NeuroEngineering and Rehabilitation*, 16(1), pp. 1–16. doi: 10.1186/s12984-019-0512-1.

Tysnes, O.-B. and Storstein, A. (2017) 'Epidemiology of Parkinson's disease', *Journal of Neural Transmission*, 124(8), pp. 901–905. doi: 10.1007/s00702-017-1686-y.

Vabalas, A. *et al.* (2019) 'Machine learning algorithm validation with a limited sample size', *PLoS ONE*, 14(11), pp. 1–20. doi: 10.1371/journal.pone.0224365.

Vale, T. C. and Cardoso, F. (2015) 'Chorea: A Journey through History', *Tremor and other hyperkinetic movements (New York, N.Y.)*, 5, pp. 1–6. doi: 10.7916/D8WM1C98.

Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods*. doi: 10.1038/s41592-019-0686-2.

Weeks, M. *et al.* (2013) 'The CARMEN software as a service infrastructure', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983). doi: 10.1098/rsta.2012.0080.

Whittle, M. W. (2007) *An Introduction to Gait Analysis*, *Library*.

Wilcox, R. (2018) 'A robust nonparametric measure of effect size based on an analog of cohen's d, plus inferences about the median of the typical difference', *Journal of Modern Applied Statistical Methods*, 17(2). doi: 10.22237/jmasm/1551905677.

Wild, E. J. and Tabrizi, S. J. (2014) *Premanifest and Early Huntington's Disease*, *Huntington's Disease*. Oxford University Press. doi: 10.1093/med/9780199929146.003.0005.

Wild, E. J. and Tabrizi, S. J. (2017) 'Therapies targeting DNA and RNA in Huntington's disease', *The Lancet Neurology*, 16(10), pp. 837–847. doi: 10.1016/S1474-4422(17)30280-6.

Winder, J. Y. *et al.* (2019) 'Longitudinal assessment of the Unified Huntington's Disease Rating Scale (UHDRS) and UHDRS–For Advanced Patients (UHDRS-FAP) in patients with late stage Huntington's disease', *European Journal of Neurology*, 26(5), pp. 780–785. doi: 10.1111/ene.13889.

Woodgate, S. *et al.* (2021) 'Objectively characterizing Huntington's disease using a novel upper limb dexterity test', *Journal of Neurology*. doi: 10.1007/s00415-020-10375-8.

Youssov, K. *et al.* (2013) 'Unified Huntington's disease rating scale for advanced patients: Validation and follow-up study', *Movement Disorders*, 28(12), pp. 1717–1723. doi: 10.1002/mds.25654.

Zampogna, A. *et al.* (2020) 'Fifteen years of wireless sensors for balance assessment in neurological disorders', *Sensors (Switzerland)*, 20(11), pp. 1–32. doi: 10.3390/s20113247.

Zhang, L. *et al.* (2019) 'Jerk as an indicator of physical exertion and fatigue', *Automation in Construction*, 104(October 2018), pp. 120–128. doi: 10.1016/j.autcon.2019.04.016.

Zhang, L., Jeong, D. and Lee, S. (2021) 'Data Quality Management in the Internet of Things', *Sensors*, 21(17), p. 5834. doi: 10.3390/s21175834.