Computer vision of video to measure bradykinesia and tremor in Parkinson’s Disease

Stefan Williams

Submitted in accordance with the requirements for the degree of Doctor of Philosophy, PhD

The University of Leeds School of Medicine

Thesis submitted for examination 30th June 2022

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and that of the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others. **The following chapters are based on work from jointly authored publications (listed, with author contributions).**

Chapter 3

Williams S, Wong D, Alty JE, Relton SD. Parkinsonian hand or clinician's eye? Finger tap Bradykinesia interrater reliability for 21 movement disorder experts. *Journal of Parkinson's Disease*. 2023 Apr 20(Preprint):1-2.

Conception: SW, JEA, SDR. Data collection: SW. Data analysis: SDR, SW. Manuscript writing / editing: SW, SDR, JEA, DW.

Chapter 4

Williams S, Relton SD, Fang H, Alty J, Qahwaji R, Graham CD, Wong DC. Supervised classification of bradykinesia in Parkinson's disease from smartphone videos. *Artificial Intelligence in Medicine*. 2020; 110:101966.

Conception: SW, HF, DCW, SDR. Data collection: SW, JA. Computer vision work: HF, SW. Data analysis: DCW, SW, SDR. Manuscript writing / editing: SW, DCW, HF, SDR, JA, CDG, RQ.

Williams S, Zhao Z, Hafeez A, Wong DC, Relton SD, Fang H, Alty JE. The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia?. *Journal of the Neurological Sciences*. 2020; 416:117003.

Conception: SW, JEA, DCW, SDR. Data collection: SW, AH, JA. Computer vision work: SW, AH. Data analysis: SW, DCW, ZZ. Manuscript writing / editing: SW, DCW, SDR, HF, JEA.

Chapter 5

Williams S, Fang H, Relton SD, Graham CD, Alty JE. Seeing the unseen: could Eulerian video magnification aid clinician detection of subclinical Parkinson's tremor?. *Journal of Clinical Neuroscience*. 2020; 81:101-4.

Conception: SW, HF, JAE, SDR. Data collection: SW, JA. Computer vision work: SW, HF. Data analysis: SDR, SW. Manuscript writing / editing: SW, JAE, SDR, HF, CDG.

Williams S, Fang H, Relton SD, Wong DC, Alam T, Alty JE. Accuracy of smartphone video for contactless measurement of hand tremor frequency. *Movement Disorders Clinical Practice*. 2021; 8(1):69-75.

Conception: SW, HF, DCW, JAE. Data collection: SW, TA, JA. Computer vision work: HF, SW. Data analysis: DCW, SW, SDR, HF. Manuscript writing / editing: SW, DCW, JAE, SDR, HF.

Acknowledgments

This thesis is the result of collaborative research. I worked on all aspects and sections of the thesis, with specific assistance, guidance and contributions from the following people, who I would like to acknowledge and thank. I am very grateful for their help and collaboration.

Dr Samuel Relton
Guidance on all aspects of the thesis. Statistical analysis. In particular, for Chapter 3, Dr Relton wrote the code for the standard linear model and cumulative linked mixed model for intraclass correlation coefficient of clinician MDS-UPDRS finger tapping scores. The use of that method to calculate ICC was his conception. For Chapter 5, tremor amplification, Dr Relton wrote the code for the mixed effects logistic regression model to calclulate the significance of the change in proportion of correctly classified hands after Eulerian magnification for three clinicians combined. Dr Relton devised Figure 3.3 in Chapter 3. Assisted with drafting the text of the publications that form the basis of Chapters 3, 4 and 5.

Dr Hui Fang
Guidance on computer vision aspects of the thesis. Computer vision code (and guidance on choice of approaches). In particular, for Chapter 4 and Chapter 5, Dr Fang wrote or adapted the code for hand region detection (Chapter 4), as well as the code to calculate histograms of optical flow for finger tapping movement (Chapter 4) and hand tremor movement (Chapter 5). He also provided guidance with using GrabCut (Chapter 4) and Eulerian magnification code (Chapter 5).

Dr David Wong
Guidance on, and assistance with, signal processing aspects of the thesis. Dr Wong wrote the code for principal component analysis, linear regression, naïve Bayes, and support vector machine models to predict finger tapping MDS-UPDRS finger tapping score category, and patient/control status, in Chapter 4. Dr Wong devised Figures 4.1 and 4.3 in Chapter 4. He also provided guidance on the choice of analysis to compare accelerometer versus computer vision measurements of hand tremor, i.e. selection of Bland-Altman analysis (Chapter 5), and guidance on performing this calculation. Assisted with drafting the text of the publications that form the basis of Chapters 3, 4 and 5.

Professor Jane Alty
Guidance on all aspects of the thesis. Assistance with drafting the text of publications that form the basis of Chapters 3, 4 and 5 (including study conception, introduction and discussion sections, the clinical context and relevance of the work).

Professor Rory O'Connor
Guidance, supervision and assistance with project and thesis overall direction, structure, writing.

Dr Marc Randall and Dr Oliver Lily
Providing flexible clinical work that enabled me to pursue this unfunded academic work.

Abstract

The assessment of Parkinson's disease is based upon clinician visual judgement. At the centre of this is a characteristic visible impairment of movement – bradykinesia – with often another visible sign, tremor. My thesis is that computer interpretation of video can provide clinically meaningful measures of finger tapping bradykinesia, and hand tremor, in Parkinson's disease.

A scoping literature review of technologies to automate the finger tapping test for bradykinesia in Parkinson's (to 2021) identified 54 studies. Published methods include surface contact, infrared, gyroscope, accelerometer. There is a wide variation in strength and significance of correlations with clinical ratings, classification accuracies and group mean differences.

Interrater reliability for judging finger tapping bradykinesia was investigated for 21 neurologists using 137 videos rated by the Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS). There was only moderate agreement, intraclass correlation coefficient 0.53 (standard linear model) and 0.65 (cumulative linked mixed model). 24% of control videos were judged as bradykinesia. 70% of videos were correctly identified as Parkinson's/control.

A computer vision optical flow method was applied to 70 finger tapping videos, with dimensionality reduction using principal component analysis before input to classification models. Test accuracy was 0.8 for mild/moderate/severe bradykinesia and 0.67 for the presence of Parkinson's disease. The computer vision pose estimation technique DeepLabCut was applied to 133 finger tapping videos. Resultant measures correlated well with clinical ratings of bradykinesia (Spearman coefficients): −0.74 speed, 0.66 amplitude, −0.65 rhythm for Modified Bradykinesia Rating Scale; −0.56 speed, 0.61 amplitude, −0.50 rhythm, −0.69 combined for MDS-UPDRS. All $p < .001$.

Eulerian video magnification was applied to 48 videos of atremulous hands. The proportion of hands correctly classified as parkinsonian/control by clinicians was higher after Eulerian magnification (OR = 2.67; CI = [1.39, 5.17]; $p < 0.003$). Optical flow with Fourier transform was applied to 40 videos of tremulous hands. Bland-Altman analysis of dominant tremor frequency from video compared with accelerometer showed excellent agreement: 95% limits of agreement –0.38 Hz to +0.35 Hz.

These results suggest that standard smartphone video can be used to derive measures of bradykinesia and tremor, and could form the basis of a tool to augment clinical assessment.

Table of Contents

Chapter 1, Introduction

In the practice of clinical medicine, visual observation plays a fundamental role, and many of the signs of disease are visible on clinical examination [1]. An experienced clinician can recognise visual patterns that are indicative of the presence or severity of a particular disease. This is particularly true for neurology, because many important neurological diseases affect human movement in a characteristic way, either reducing movement, changing movement, or adding new movement [2, 3]. Most neurological diseases do not have simple laboratory or radiological tests, so that clinician assessment, including clinician visual assessment, is often at the centre of diagnosis and the grading of severity [4-10]. An archetypal example of this is the neurodegenerative movement disorder Parkinson's disease. It was first described in largely visual terms [11], and the modern criteria for diagnosis [7] and severity rating [8] both rely on the clinician's visual judgement during clinical examination.

A reliance on clinician visual judgement places limits on clinical assessment, because human visual judgement is inherently limited in its accuracy for detecting and measuring subtle differences in movement. In recognition of this, there has been a long history of published reports of technology to assess movement in neurological disease [12-47]. However, these have involved either a requirement for new hardware or a requirement for patients to engage in a new behaviour by using a smartphone app or website. No such technology has entered routine clinical practice. In contrast, computer interpretation of video – a form of computer vision – could in principle provide automated or objective assessment of the visual signs of neurological disease. That would not require new equipment or new patient behaviour, since clinicians already observe patient movement during examination, and digital cameras are ubiquitous. Two of the cardinal clinical signs of Parkinson's disease can be observed in the hands: bradykinesia and tremor. My thesis applies computer vision to smartphone video of the hands of people with Parkinson's disease.

Parkinson's disease

Parkinson's disease is a progressive, neurodegenerative condition. The ultimate 'gold standard' for the presence of the disease is brain pathology found at postmortem [48]. In the early 20th century, the pathological hallmark of the disease, intraneuronal inclusions termed Lewy bodies, were recognised, including within substantia nigra neurons in the midbrain [49]. Lewy bodies contain aggregates of the protein alpha-synuclein, and current understanding of pathophysiology is that there is a complex interplay between alpha-synuclein aggregation, inflammation, and cell storage and transport, that results in accelerated neuronal death of primarily dopaminergic neurons but also multiple other neuronal pathways [49, 50]. This is thought to lead to an imbalance between basal ganglia pathways in the brain that inhibit versus facilitate movement [51-53]. The 'Braak hypothesis', based on postmortem work, suggests that Parkinson's pathology begins in the gut and ascends rostrally over time, through the brainstem, then the basal ganglia (causing movement symptoms) and finally affecting the cortex (associated with cognitive impairment) [54].

The pathological changes in the brain in Parkinson's can only be tested for after death, in the minority of people who undergo postmortem. As such, in life, the diagnosis of Parkinson's disease is made by clinical assessment. The current clinical definition of the condition is contained in the 2015 Movement Disorder Society diagnostic criteria [7]. Within this, the fundamental requirement is the presence of 'parkinsonism' on examination. At the core of parkinsonism is a movement abnormality termed bradykinesia, because parkinsonism is defined as limb *bradykinesia*, plus one or both of *resting tremor* or *rigidity* [7]. Bradykinesia is defined as:

"slowness of movement AND decrement in amplitude or speed (or progressive hesitations/halts) as movements are continued" [7].

Bradykinesia is detected by observation of specific repetitive limb movements, which the diagnostic criteria state "can be" finger tapping, hand open and close movements, upper limb pronation-supination movements, toe tapping, and foot tapping [7]. If parkinsonism is established, then Parkinson's disease is diagnosed when that parkinsonism is combined with the presence of supportive criteria, such as response to dopaminergic medication, as well as the absence of exclusion criteria, such as cerebellar abnormalities, and balanced against any ‘red flags’, such as rapid gait impairment [7].

Once Parkinson’s is diagnosed, then at each clinical review the clinician must judge the severity of the condition. Research trials also require assessment of disease severity, both for participant selection and outcome measures. The current standard research rating scale for the severity of Parkinson’s is the Movement Disorder Society sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS) [8]. This has four parts: Part I (non-motor experiences of daily living), Part II (motor experiences of daily living), Part III (motor examination) and Part IV (motor complications). Part III, the motor examination, is almost entirely based on visual observations, each of which the clinical examiner scores on a scale from 0 (normal) to 4 (severe impairment) [8]. It includes limb bradykinesia items, such as repetitive finger-thumb tapping and hand pronation-supination.

In busy clinical practice, there is not enough time to apply the full MDS-UPDRS (described as “a vast instrument” [55]). Nevertheless, the shorter, more focused examination in routine care is based on the same principles as the MDS-UPDRS. Since bradykinesia is at the centre of the clinical concept of Parkinson’s, a clinical assessment almost always involves a judgement of bradykinesia severity. One of the most common methods to ascertain the presence and severity of bradykinesia is finger tapping, whereby the neurologist observes the patient repeatedly tapping their index finger against thumb “as quickly and as big as possible” [8]. This is item 3.4 in the motor examination section of the MDS-UPDRS. In that scale, three elements of finger tapping bradykinesia - speed, amplitude and rhythm - are assessed into a composite score between 0 and 4, **Table 1.1**. An MDS-UPDRS finger tapping score above 0 does not necessarily mean bradykinesia is present, since any single element of bradykinesia in isolation will raise the MDS-UPDRS score above 0, without meeting the definition of bradykinesia. The MDS-UPDRS finger tapping score thus measures *severity*, but not necessarily *presence*, of bradykinesia. In contrast, a more recent rating scale, the Modified Bradykinesia Rating Scale (MBRS), rates each bradykinesia component separately, and includes a finger tapping item, **Table 1.2** [27, 56]. Subscores from that scale can thus also indicate the *presence* of bradykinesia, in addition to the *severity*.

Score	Criteria
0: Normal	No problems
1: Slight	Any of the following: a) the regular rhythm is broken with one or two interruptions or hesitations of the tapping movement; b) slight slowing; c) the amplitude decrements near the end of the 10 taps.
2: Mild	Any of the following: a) 3 to 5 interruptions during tapping; b) mild slowing; c) the amplitude decrements midway in the 10-tap sequence.
3: Moderate	Any of the following: a) more than 5 interruptions during tapping or at least one longer arrest (freeze) in ongoing movement; b) moderate slowing; c) the amplitude decrements starting after the 1st tap.
4: Severe	Cannot or can only barely perform the task because of slowing, interruptions or decrements.

Table 1.1. Item 3.4, finger tapping, from the MDS-UPDRS (Movement Disorders Society sponsored revision of the Universal Parkinson's Disease Rating Scale). Each hand is tested separately. The patient is instructed to tap the index finger on the thumb 10 times "as quickly AND as big as possible" [8].

Score	**Speed**	**Amplitude**	**Rhythm**
0	Normal	Normal	Regular, no arrests or pauses in ongoing movement
1	Mild slowing	Mild reduction in amplitude in later performance, most movements close to normal	Mild impairment, up to two brief arrests / 10 seconds, none lasting > 1 second
2	Moderate slowing	Moderate reduction in amplitude visible early in performance but continues to maintain 50% amplitude through most of the task	Moderate, 3 to 4 arrests / 10 seconds; or 1 or 2 lasting > 1 second
3	Severe slowing	Severe, less than 50% amplitude through most of the task	Severe, 5 or more arrests / 10 seconds; or more than 2 lasting > 1 second
4	Can barely perform the task	Can barely perform the task	Can barely perform the task

Table 1.2. The Modified Bradykinesia Rating Scale. Each hand is tested separately. The patient is instructed to tap the index finger on the thumb "as quickly AND as big as possible" for 10 seconds [27].

The motor examination section of the MDS-UPDRS also includes other visible impairments of movement such as reduced ("masked") facial expression (hypomimia) and impaired gait (including impairments in stride amplitude, stride speed, arm swing). Furthermore, it is important to note that the overall MDS-UPDRS is not entirely based on visible movement problems. Part 1 of the MDS-UPDRS (non-motor experiences of daily living) recognises that Parkinson's also causes a variety of non-motor symptoms, that can sometimes cause more suffering than the motor symptoms. These include sleep-wake dysregulation, mood disturbance, cognitive dysfunction, urinary symptoms, and constipation.

There is considerable individual variation in both the clinical features of idiopathic Parkinson's and also the prognosis [57]. It is likely that idiopathic Parkinson's represents several subtypes of disease and there is some evidence for division into 'akinetic-rigid' versus 'tremor-predominant' subtypes, the latter with a slower progression [58]. There appears to be a very long prodromal phase of Parkinson's disease that begins many years prior to the onset of motor symptoms, perhaps even 20 years before [59-61]. Symptoms such as REM-sleep behaviour disorder, anosmia, constipation and depression can precede the development of parkinsonism by many years. Such factors have varying sensitivity and specificity for the future development of Parkinson's, and are strongest predictors in combination.

Parkinson's disease is most commonly idiopathic (sporadic). In 2016, there were 6.1 million people worldwide with the disease [62]. Some lifestyle or environmental factors are associated with increased risk of idiopathic Parkinson's, such as pesticide exposure, consumption of dairy products, and head injury [63]. There are also factors that are associated with lower risk such as physical exercise, smoking and coffee consumption [63]. However, none are strong associations and causality has not been demonstrated. Genetic variants play a role in idiopathic Parkinson's but there are multiple loci, together explaining only a minority of the risk [64, 65]. Monogenetic Parkinson's disease exists but is rare (e.g. PARK mutations) [65]. In contrast to idiopathic Parkinson's disease, parkinsonism can also be drug-induced (most commonly antipsychotic, dopamine blocking drugs), secondary to vascular disease, or due to sporadic 'Parkinson's plus' neurodegenerative conditions that are rarer than idiopathic Parkinson's and have distinct clinical features, such as autonomic failure in the condition multiple system atrophy [57].

The historical development of the concept of Parkinson's disease

This current clinical conception of Parkinson's disease is one that has developed and evolved over time. Although there are apparent descriptions of parkinsonism in earlier literature [66, 67], the beginning of the modern idea of the disease is James Parkinson's 1817 'Essay on the Shaking Palsy' [11]. In this descriptive study of six people, Parkinson did not use the term bradykinesia, or describe the specific elements in the modern definition of bradykinesia. Instead, he referred to a “lessened muscular power” (“palsy”) [11]. However, he also described a range of other clinical features, all of which relate to those that we know today. This includes tremor, festinant (shuffling) gait, speech disturbance, and non-motor symptoms such as sleep disruption and constipation. In addition, he recognised the progressive, degenerative nature of the condition.

In the 1870s, Jean-Martin Charcot first named the disease after Parkinson, and noted muscular rigidity (a non-visual sign of the disease). He crucially made the observation that the sufferers were not paralysed, and that “the muscular power is retained, to a great degree” [68]. Rather, he observed “in the execution of movements, a notable want of ease”, and that “their problem relates more to slowness in execution of movement rather than to real weakness” [68]. This is different from the modern concept of bradykinesia, with its specific, multifaceted definition. Charcot did *not* specify a combination of slowness with decrement or progressive hesitations and halts, during repetitive movements.

The term bradykinesia appears in literature from 1907, but was considered descriptive of a number of different diseases, not just Parkinson’s [69]. It was used to describe a combination of features of which slowness was one, without other elements of the modern definition. From the 1920s onwards the terms bradykinesia, hypokinesia and akinesia appeared in textbooks in relation to Parkinson’s [69]. The Greek translations are slow (brady-), decreased or smaller (hypo-), no (a-), and movement (-kinesis). However, those terms have since been used to describe broader ranges of movement abnormalities than their literal translations, and used inconsistently and interchangeably [69]. So, it is notable that the current concept of bradykinesia – at the centre of the current clinical definition of Parkinson’s – does not have a long and stable history.

In the 1960s there was a gradual discovery that replacement of dopamine, in the form of oral levodopa, could significantly ameliorate many of the symptoms of Parkinson’s disease, including motor symptoms [67, 70]. Dopaminergic therapy remains the mainstay of Parkinson’s treatment today [57]. Although dopamine-based treatment (and sometimes deep brain surgery) can greatly ameliorate symptoms, there are as yet no known agents to slow progression of the underlying neurodegeneration [57].

The advent of therapy led to a requirement to formulate specific research criteria for diagnosis (trial inclusion) and scales for severity (trial outcome) [69]. The first diagnostic criteria were introduced in 1988 ('UK brain bank criteria'), in which the term bradykinesia was used, to specify slowness of movement plus progressive reduction in speed and amplitude of repetitive actions [71]. This became the closely related current definition in the 2015 diagnostic criteria. To gauge the severity of Parkinson's, the Unified Parkinson's Disease Rating Scale (UPDRS) was introduced in 1987 [72], revised in 2008 (MDS-UPDRS) [8], and has remained the standard research rating scale.

Tremor

In parkinsonism, and so Parkinson's disease, there are three core clinical signs. Bradykinesia – a visual sign – is at the centre of these. However, of the other two signs, rigidity and tremor, it is notable that tremor is also a visible sign. Tremor can be defined as an involuntary, rhythmic, oscillatory movement of a body part [73]. In routine clinical practice, the presence, pattern and severity of tremor is judged by the eye of the examining clinician.

Most, though not all, people with Parkinson's have a tremor, but there are many other conditions that can cause tremor. The International Movement Disorder Society Consensus Statement on the Classification of Tremors uses two axes. Axis 1 classifies according to clinical features, **Figure 1.1**. Axis 2 classifies the aetiology of the tremor (genetic, acquired or idiopathic). Combinations of Axis 1 clinical features correspond to specific syndromes (diagnoses), shown in **Figure 1.2**. Identification of a specific syndrome can aid the identification of aetiology (Axis 2), but classification can change over time as a disease progresses, and just as a clinical syndrome can have multiple aetiologies, a single aetiology can produce more than one syndrome.

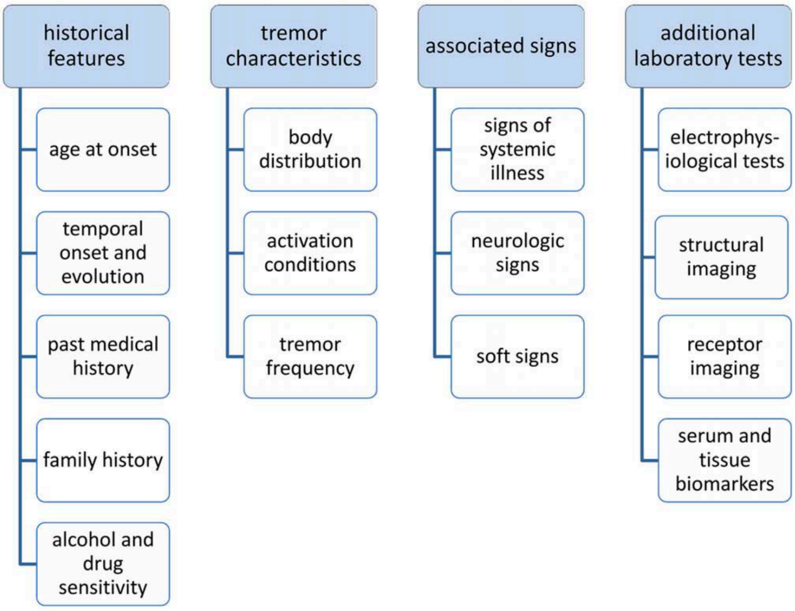


Figure 1.1. Axis 1: Clinical Features, from the Movement Disorders Society consensus statement on the classification of tremors [73]

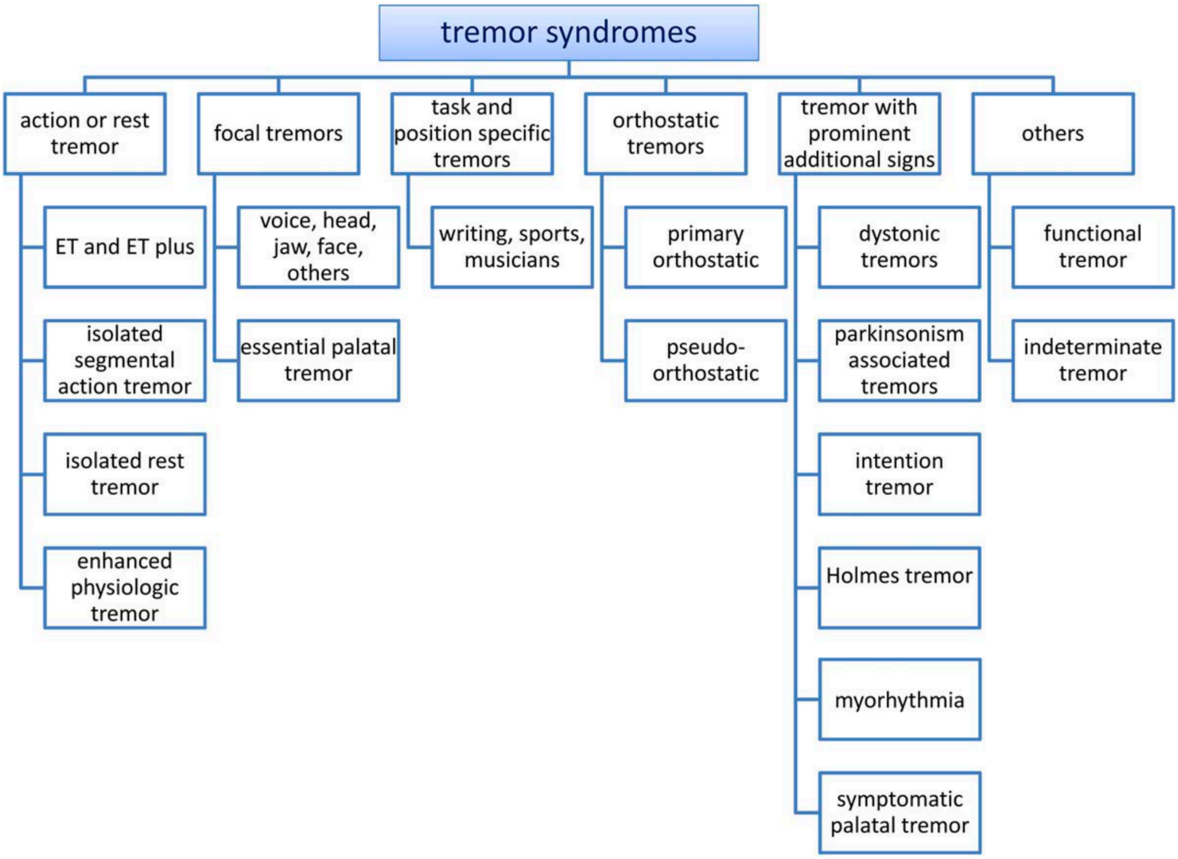


Figure 1.2. Tremor syndromes that represent particular combinations of Axis 1 clinical features, from the Movement Disorders Society consensus statement on the classification of tremors [73]

Several of the features within Axis 1 are visual signs, detected by clinician observation during examination [74]. This includes body distribution and activation conditions, such as changes in tremor severity with rest, posture or movement of the affected body part, as well as tremor alteration during examination routines such as contralateral hand tapping. Additional neurological signs are also frequently visual, such as the presence of bradykinesia in Parkinson's tremor. In addition, tremor frequency is often mentioned within descriptions of the clinical examination features for a particular tremor syndrome, implying that a clinician might attempt to estimate the approximate frequency of the tremor during visual assessment [74, 75].

The classic tremor of idiopathic Parkinson's disease is described as occurring at rest, at a frequency of 4 to 7 Hz, often asymmetrically and mainly involving the distal limbs (as well as sometimes jaw or tongue) [73, 75, 76]. A pill-rolling tremor is sometimes seen, involving thumb flexion from a neutral position, which is considered specific to Parkinson's and some forms of parkinsonism [75]. However, in addition to the classic description, other forms of tremor also occur in Parkinson's. In particular, a 'reemergent' postural hand tremor is common, which is viewed as a resting tremor that re-emerges after stable posturing (i.e. hands outstretched), suggesting that Parkinson's tremor is a 'tremor of stability' [76, 77]. Furthermore, a different postural tremor is also found in some Parkinson's patients, which has a higher frequency and begins immediately (without reemergence) [76]. It should also be noted that tremor is not universal in Parkinson's, such that some patients never develop it [58, 78].

The most common tremor disorder is not Parkinson's disease, it is a condition termed Essential Tremor (ET) [79]. Essential tremor is frequently on the differential diagnosis list for early Parkinson's tremor [74, 80-82]. Essential Tremor is defined as an isolated syndrome of bilateral upper limb action tremor, for at least 3 years duration, with or without tremor in other locations, and in the absence of other neurological signs [73]. However, Essential Tremor Plus is also recognised, in which there are "additional neurological signs of uncertain significance", including impaired tandem gait and memory impairment [73]. This 'plus' concept is controversial, and it is not based on any clear difference in pathology or prognosis [83, 84]. The frequency of Essential Tremor is often described as 4-12 Hz, a broad range, but the frequency shows an inverse correlation with age [85, 86]. Although Essential tremor is defined as a postural tremor, it can occur at rest in advanced disease [85] and rest-tremor is included among ET-plus features [73].

Other common tremor syndromes include functional (psychogenic) tremor, "characterised by distractibility, frequency entrainment or antagonistic muscle coactivation", and dystonic tremor, in which tremor and dystonia (sustained involuntary contractions of agonist and antagonist muscles) are combined as the main neurological signs [73].

The pathophysiological mechanisms of tremor disorders are not well understood. The 'dimmer-switch' model of Parkinson's tremor suggests that the basal ganglia activate tremor (operating as a 'light switch'), while a cerebello-thalamo-cortical circuit modulates tremor amplitude (analogous to a 'light dimmer'). In Essential Tremor, there is increased activity in a cortico-olivo-cerebello-thalamic circuit that is coherent with tremor oscillation of muscle activity, but the driver of this circuit is unclear, possibly related to degeneration in the cerebellum [87].

Electrophysiological measurement can complement the examination and history in the clinical assessment of tremor. Both accelerometer and gyroscope devices have been used to measure the movement of tremor, commonly attached to the dorsum of the hand [86, 88]. Movement in space has 6 degrees of freedom (3 translational, 3 rotational). Although most tremors involve primarily rotatory movement about a joint, until recently gyroscopes were too large and heavy for tremor measurement [88] and so accelerometers have been more widely used [86, 88]. An accelerometer produces a time series of linear acceleration in units of cm/s^2 or g, which can then be converted into the frequency domain by Fast Fourier Transform, allowing determination of dominant tremor frequency. A single (one) axis accelerometer is more affordable in clinical settings, and produces a single time series rather than six: a simpler signal for clinician analysis [86]. The accelerometer is attached to the patient in such a way that the single axis is aligned with the main the main axis of the tremor movement. An accelerometer measurement of tremor frequency can be combined with electromyography (EMG) to record the contraction of individual agonist and antagonist muscles participating in the movement [86].

Tremor electrophysiology can assist tremor classification, for example by demonstrating the variable frequency over time that is characteristic of the diagnosis of Functional Tremor [1]. However, such techniques represent a limited and time-consuming resource, usually reserved for a small subset of patients in specialist hospital settings. Furthermore, accelerometers are subject to gravitational artefact [88]. Whilst smartphone accelerometers can measure the dominant frequency of hand tremor, the patient must hold the phone or have it strapped in place [89]; and this will potentially alter tremor characteristics, for example 'weight loading' can alter tremor characteristics [86].

The visual judgement of Parkinson's and tremor conditions is difficult

A reliance on clinician judgement by eye for diagnosis and assessment has two problems. The first is that it requires an experienced clinician (a limited resource). The second is that humans cannot accurately measure movement by eye, or easily detect small, subtle changes. That is perhaps particularly true for Parkinson's and related conditions, because the current definition of bradykinesia describes a complex and heterogenous phenomenon. Furthermore, like those for Parkinson's, the diagnostic criteria for Essential tremor have varied substantially in the past, suggesting that the most common clinical tremor syndrome contains significant ambiguity [90]. This is likely to fundamentally limit accuracy of clinical decisions, both for diagnosis and monitoring, and research outcomes.

Consistent with this, a study reviewing 71 consecutive patients with a pre-evaluation diagnosis of essential tremor found that 37% were misdiagnosed, with the most common true diagnosis being Parkinson's disease [91]. Furthermore, when shown video recordings of the neurological examination for patients with a variety of tremor disorders, two movement disorder specialist neurologists showed high false positive (17.4-26.1%) and negative (6.7-20%) rates for the diagnosis of Parkinson's [92]. Pathology studies suggest the validity of diagnosis of Parkinson's disease in life is far from perfect. One study reported sensitivity of

88%, specificity of 66% [93], and a meta-analysis gave diagnostic accuracy figures of 73.8% for non-experts and 79.6% for movement disorder specialists [94].

Severity rating scales (based on visual judgement) are also limited in their accuracy. In one study, 226 raters were tested in their UPDRS motor (examination) scores for 4 people with Parkinson’s disease using video recordings [95]. A ‘pass’ in this test was defined as a score within the 95% confidence interval of 3 “international Parkinson’s disease experts” for each case. Only 54.6% of raters ‘passed’ the 4 cases. In another study, three trained nurses and one movement disorder specialist rated older people *with no clinical Parkinson’s*, according to a modified UPDRS motor score [96]. They gave 74 out of 75 participants a score greater than 0 (mean score 13.4 out of 127).

Computer vision

Given the limitations of human judgement for Parkinson’s and tremor disorders, there have been many published attempts to use technology to automate measurement of the clinical signs of Parkinson’s, including bradykinesia and tremor [12-47, 80, 97]. One approach is to use specific equipment (often ‘wearable’) that can measure movement. For bradykinesia, examples include devices based on accelerometers, gyroscopes, and infrared cameras (usually with active or passive markers attached to the body) [12-47]. For tremor, measurement in a laboratory setting uses accelerometer and electromyography [86, 98], and portable accelerometer devices can record tremor in non-controlled environments such as a standard clinic room [88].

The other main approach has been to use smartphones to record limb bradykinesia or tremor. For bradykinesia, this most commonly involves a person with Parkinson’s tapping the phone screen, or sometimes strapping the phone to the distal portion of a limb [22, 46, 99, 100]. For limb tremor, studies have used the smartphone’s inbuilt accelerometer, with the patient holding the phone in their hand or strapping it in place [80, 97].

None of these methods have entered routine clinical practice for the assessment of patients. The specialist 'wearable' or hardware approach is limited by a requirement for large scale purchase and maintenance of specific devices. The smartphone screen tapping or smartphone accelerometer method is limited in the data that can be acquired. Tapping a screen cannot measure movement amplitude – a fundamental part of the concept of bradykinesia [7]. Holding a phone adds weight to a limb, which alters tremor characteristics, a phenomenon known as 'weight-loading' [86]. Most importantly, these smartphone methods require patient motivation to use and interact with a specific app, potentially on more than one occasion, a change in behaviour that is perhaps unlikely to become widespread.

There is a potential alternative to these previous approaches. That is to use two ubiquitous items of equipment – the camera and the computer – to capture and assess the information that is seen by clinicians during the neurological examinations that they already perform. Computer vision can be defined as the ability of a computer to make "useful decisions about real physical objects and scenes based on sensed images" [101]. 'Sensed images' most commonly refers to input from standard monocular cameras or video cameras [102]. Computer vision of video has developed to a high level of accuracy in non-medical applications. For example, autonomous steering and lane detection in cars [103, 104], traffic monitoring [105], crowd analysis [106], facial recognition [107]. Computer vision does not require special equipment or patient motivation, and is contactless (relevant in the light of COVID-19).

Computer vision is potentially well-suited to detecting and measuring the visual signs of neurological disease. There are many published algorithms that can, in general, recognise and track objects, people and movement in video. These are broad enough in their mechanisms that in principle they could be applied to video of the neurological examination. Three techniques that can be applied to human movement in video are optical flow, pose estimation and Eulerian magnification.

Optical flow

Optical flow can be defined as "the distribution of apparent velocities of movement of brightness patterns in an image" [108]. It arises from the relative motion of objects and the viewer. If a camera is static, then movement of an object (such as a person or a body part) within the camera frame creates a moving brightness pattern. Computer vision algorithms can detect and measure optical flow in video, and this provides a method to record movement. Most commonly, video involves opaque objects of finite size undergoing rigid motion or deformation. This means neighbouring points on an object have similar velocities, so that the velocity of brightness patterns in the image varies smoothly almost everywhere. Optical flow algorithms relate the change in image brightness at a point to the motion of the brightness pattern. These techniques convert video frames into an optical flow field [108]. In such a field, each position corresponds to the vector pixel movement of a point object between two sequential video frames. The magnitude of the vector represents the instantaneous speed of a point (in pixels/frame). Optical flow vectors can be divided into bins corresponding to movement direction, which can then be used to produce histograms of optical flow, showing the pattern of movement direction across pairs of video frame.

Figure 1.3 shows an illustration of optical flow vectors derived from video of a walking sequence, taken from reference [109]. **Figure 1.4** shows illustrates the technique of dividing optical flow vectors into bins corresponding to movement direction, while **Figure 1.5** shows the resultant histograms of optical flow after dividing the walking sequence into direction bins.

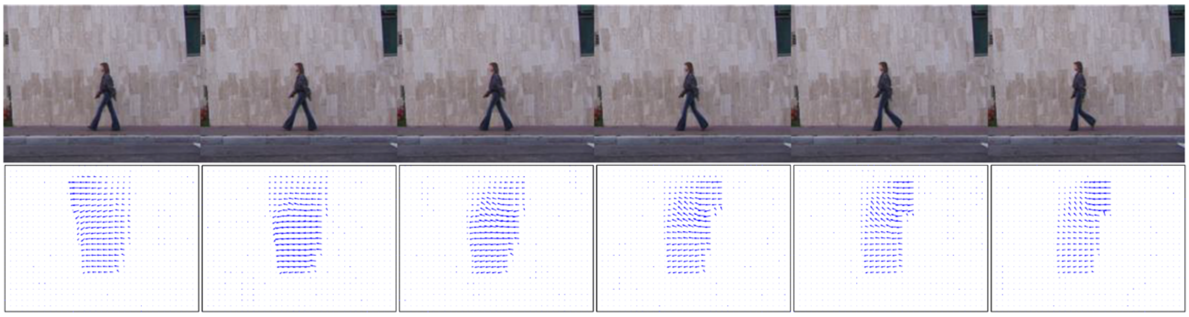


Figure 1.3. Optical flow patterns for a simple walking sequence. The blue arrows show the directions and magnitude of pixel movement for the corresponding video frames. Taken from reference [109]

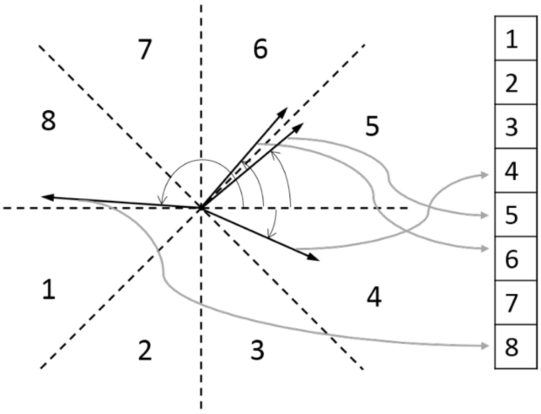


Figure 1.4. An illustration of the Histogram of Optical Flow (HOOF) method whereby pixel optical flow vectors are divided into 8 bins of movement direction (boxes 1 to 8). Taken from reference [110]

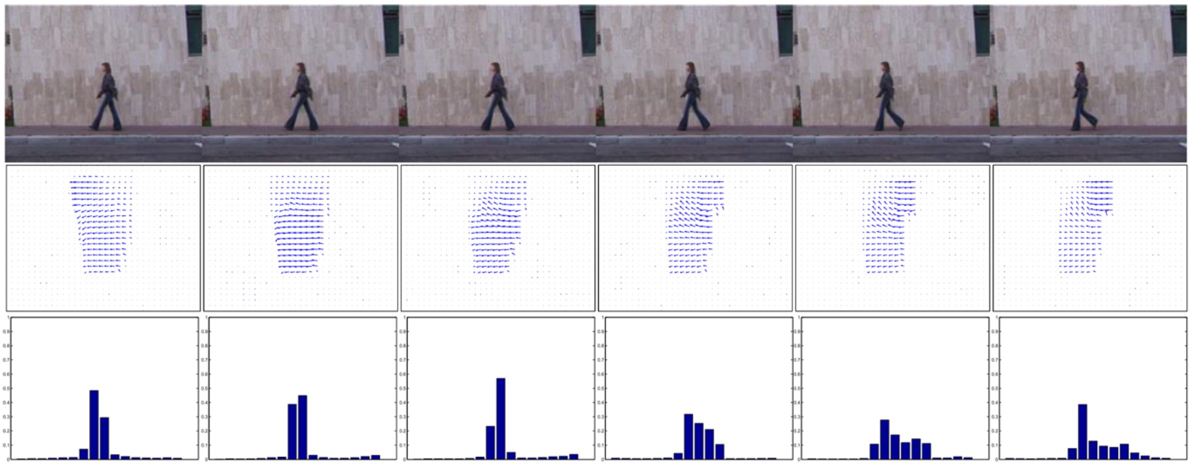


Figure 1.5. An example of optical flow patterns (second row of images) converted into Histograms of Optical Flow (third row of images) for a simple walking sequence. Taken from reference [109]

Pose estimation

'Pose estimation' refers to computer vision algorithms that estimate the location of human body parts and how they are connected in an image [111]. Most commonly, this relates to video from a standard monocular camera. There are multiple applications, for situations in which the automatic tracking of human movement and action in video might provide useful information. Pose estimation has traditionally involved two phases. The first is to detect body parts or estimate their position [112]. The second is an inference step that combines local observations of body parts with spatial constraints. Originally, part detectors used a variety of different machine learning classification techniques, but the most modern and most effective pose estimation algorithms all use deep learning techniques to detect body parts. Deep learning does not require explicit features specified by researchers, and instead learns the most predictive features of a particular category directly from images, given a large data set of labelled examples. Deep learning involves a process termed back-propagation that indicates how the algorithm should change its internal parameters to best predict the desired output of an image [113, 114]. Current pose-estimation algorithms are trained on large image datasets, such as ImageNet [115]. As deep learning detection of body part positions has improved, the need for an inference step with spatial constraints (i.e. rules to connect the body parts) has lessened [112].

DeepLabCut is a pose estimation algorithm that can track the body parts of animals in video without the need for a large labelled dataset [116, 117]. It was devised for laboratory experiments involving animal behaviour models. DeepLabCut does not require a large labelled dataset of the specific video that it is applied to, because it uses transfer learning. Transfer learning is the ability to take a network that was trained on a task with a large supervised data set, and utilise it for another task with a small supervised data set. DeepLabCut uses the feature detector from a pose estimation algorithm named DeeperCut [112], that has been pre-trained on ImageNet [115], a massive object recognition dataset.

Eulerian magnification

There is a third computer vision technique that has potential relevance to video of Parkinson's and other tremor conditions. It is termed Eulerian video magnification. This is a method that is applied to standard video, to produce a new video in which the size (amplitude) of movement has been increased [118]. This means that very low amplitude movement, which was difficult or impossible to see by eye in the original video, becomes visible in the video after Eulerian magnification. Thus, it can reveal low amplitude movement in video, that was otherwise unseen. Eulerian magnification exaggerates motion by amplifying temporal colour changes at fixed positions. The variation of pixel values over time is amplified in a spatial multiscale manner. The technique first decomposes the video into different spatial frequency bands, and then performs temporal processing on each spatial band. A bandpass filter is used to isolate a temporal frequency band of interest, so that the technique amplifies select temporal frequencies of movement.

Techniques to analyse computer vision data

Optical flow and pose estimation process video so that human movement is converted to video pixel coordinate changes over time. That is time series data: a graph of time on the x axis and a measure of change in video pixel coordinates on the y axis. To obtain meaningful summary measures of neurological examination signs, such as tremor or bradykinesia, further processing of that time series data is required.

The movement of the finger tapping test for bradykinesia can potentially be tracked in video by optical flow or pose estimation techniques. A number of different measurements from this data could be used to produce a time series. One possibility is that pose estimation could track the relative position of finger and thumb tip, and relative distance between them, over time. From this raw time series, features can be chosen, that are judged to reflect the clinical construct of bradykinesia. The selection of features in this way has previously been applied to time series data derived from gyroscope (non-video) recordings of finger tapping, with gyroscope attached to finger tip. For example, angular velocity has been used as a measure of tapping speed, while coefficient of variation of the sliding window of the excursion angle was used as a measure of tapping rhythm [27]. Each of these measures can be correlated with clinician measures of bradykinesia as a way of testing convergent validity.

In addition to deriving single measurements from a time series, it is also possible to combine multiple individual features from a time series using the technique of principal component analysis (PCA). This reduces the dimensionality of a dataset whilst preserving as much of the variance in the data as possible [119]. PCA achieves this by finding new variables that are linear functions of those in the original dataset, which successively maximise variance (the principal components). It is a way to find patterns in data that has many dimensions (features)

[120, 121], because high dimensional data cannot be graphically represented to find patterns (for example, an 8-dimensional graph cannot be plotted).

Once principal components have been derived from features of a finger tapping time series, it is possible to apply machine learning classification techniques. Machine learning models can 'learn' (or be 'trained') to classify data into categories. For example, machine learning methods could be applied to principal component data to distinguish Parkinson's tapping from control tapping, in other words how well finger tapping data can be used to classify tapping as Parkinson's or not. There are several methods for this type of classification task. Logistic regression finds a classification boundary that provides the best linear separation between two categories [122]. A Naïve Bayes classifier calculates the probability of a given classification, using Bayes theorem (with the assumption that all attributes of a data point are independent of each other) [123]. A Support Vector Machine is a classification algorithm that predicts an optimal hyperplane between two categories in n-dimensional space, with a maximal margin between the classified data points (created by data points closer to the hyperplane: support vectors) [122].

Tremor is an oscillatory movement, and the frequency or frequencies of a tremor can help to characterise different underlying neurological conditions. Tremor frequency is part of 'Axis 1' of the Movement Disorders Society consensus statement on the classification of tremor, **Figure 1.1** [73]. A method to summarise the frequency of a time series signal is the Fourier transform [124], which converts a signal in the time domain to a signal in the frequency domain. In his 1822 work *The Analytical Theory of Heat*, Fourier demonstrated that arbitrary functions could be written as a summation of sines and cosines [125]. The Fourier transform converts a time series signal into a series of sine waves, each with its own amplitude and frequency. This allows determination of the relative contributions of different frequencies to a given time signal. This is a method by which dominant tremor frequency could be determined from a measure of oscillatory movement in video. Fourier transform to detect tremor frequency has previously been applied to (non-video) accelerometer data [126]. **Figure 1.6** illustrates the way in which a complex signal is the sum of individual sine waves. **Figure 1.7** illustrates the way in which the Fourier transform converts a complex time signal into the frequency domain.

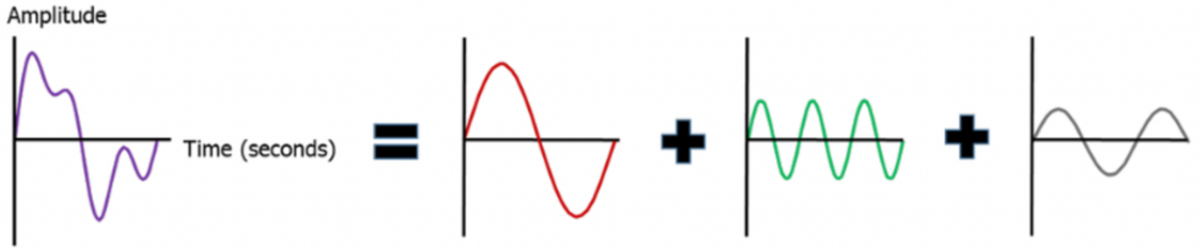


Figure 1.6, A complex time signal is the sum of individual sine waves. Taken from reference [127].

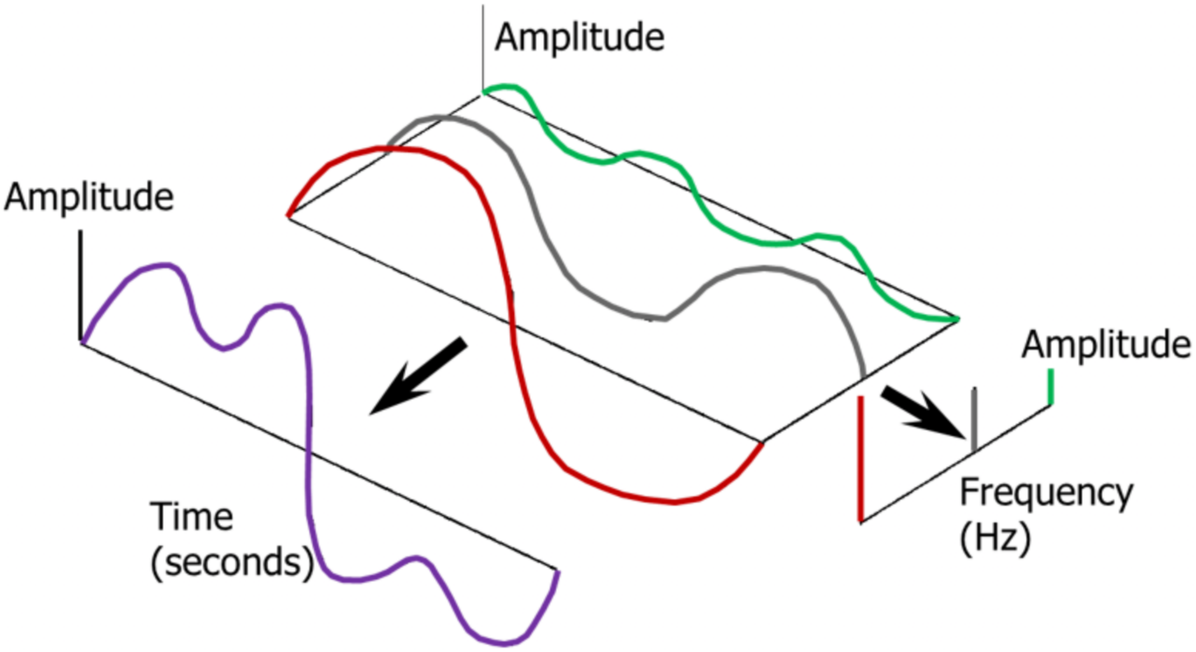


Figure 1.7. A graphical illustration of the principle by which the Fourier transform can convert a signal from the time domain to the frequency domain. Taken from reference [127].

In comparing two measurement techniques for the same measure, such as video versus accelerometer for tremor frequency, a Bland-Altman analysis is more appropriate than correlation. Correlation is not a measure of agreement; it measures the strength of relationship between two variables (which can be strongly related but not in agreement). Bland-Altman calculates the bias (mean difference) between two measurement methods, d; the standard deviation of the differences, s; and the 95% confidence intervals for the mean difference, which are referred to as ‘limits of agreement’ [128-130].

Applying computer vision to video of bradykinesia and tremor

My thesis is that techniques for computer interpretation of video can provide clinically meaningful measures of finger tapping bradykinesia, and hand tremor, in Parkinson's disease. Chapter 2 describes the literature on existing technological methods to measure finger tapping bradykinesia. Chapter 3 investigates the current 'gold standard' for bradykinesia assessment, namely clinician visual judgement. The study reports the intraclass correlation coefficient for judgement of finger tapping bradykinesia severity, in a large group of expert clinicians. It also reports the rates of apparent bradykinesia presence in Parkinson's and healthy controls, as well as the accuracy of clinicians' blinded judgement of hands as parkinsonian or not. Chapter 4 reports two new computer vision methods to measure finger tapping bradykinesia in video. It describes a measure of hand movement during finger tapping that is derived from optical flow. Features from this are combined with machine learning techniques to discriminate low from high bradykinesia, and Parkinson's from controls. It also describes a method using DeepLabCut to track the distance from finger tip to thumb tip during finger tapping, and test the correlation between measures derived from that and clinician ratings of bradykinesia severity. Chapter 5 reports two applications of computer vision of video to hand tremor. The first is the application of Eulerian video magnification to videos of hands that show no visible tremor, in both people with Parkinson's and healthy controls. The ability of clinicians to correctly identify hands as parkinsonian or not is tested before and after Eulerian magnification. The second uses an optical flow method to record the oscillatory movement of visible hand tremor in video, and derive the dominant frequency of the tremor by Fourier transform. This frequency is compared with that obtained by accelerometer for the same hands, and agreement between video and accelerometer methods are tested by a Bland-Altman comparison. Chapter 6 presents an overall discussion and conclusions.

Chapter 2, A review of technology to measure finger tapping bradykinesia in Parkinson's

Introduction

The current understanding of Parkinson's disease places one clinical feature above all others: the abnormality of movement known as bradykinesia. The 2015 International Movement Disorder Society definition of bradykinesia is:

"slowness of movement AND decrement in amplitude or speed (or progressive hesitations/halts) as movements are continued" [7].

In order to diagnose Parkinson's disease, 'parkinsonism' must be present, which refers to bradykinesia plus rest tremor or rigidity [7]. Furthermore, at subsequent clinical reviews over the course of the disease, the clinician will invariably examine for the severity of bradykinesia, to assess progression and treatment response. As described in Chapter 1, the modern concept of bradykinesia is a relatively recent development, largely dating from the 1980s, and driven by a need to definite clinical features of Parkinson's in the era of drug trials [69].

The detection and assessment of bradykinesia are clinical tasks, in which a clinician makes a visual judgement of the patient's movement [7, 8]. In current practice (and diagnostic guidelines or severity scales), the patient is asked to make one or more of the following repetitive limb movements: finger tapping, hand opening and closing, forearm pronation-supination, toe tapping, or foot tapping. These five movements are specified within the standard research tool for grading the severity of Parkinson's: the Unified Parkinson's Disease Rating Scale (UPDRS) [72], and its 2008 Movement Disorder Society sponsored revision (MDS-UPDRS) [8]. The clinical rater is asked to score the severity of bradykinesia from 0 to 4, **Table 1.1**.

The MDS-UPDRS bradykinesia rating combines deficits in movement speed, amplitude and rhythm into a single composite score. This means that the same score can result from more than one pattern of impairment, so that some individual differences in movement impairment are lost as they are grouped into the same score. In recognition of this, a newer bradykinesia rating scale was published in 2007, the Modified Bradykinesia Rating Scale (MBRS) [27, 56]. This requires the rater to separately grade movement speed, amplitude and rhythm, **Table 1.2**.

In busy clinical practice, there is rarely time to examine the five different repetitive movements used to assess bradykinesia (finger tapping, hand open/close, forearm pronation/supination, toe tapping, and foot tapping). In the clinical experience of this author, the most common single examination for bradykinesia is the finger tapping test, carried out in nearly every clinical assessment of a person with Parkinson's. The finger tapping test is thus a fundamental part of Parkinson's assessment: an almost universally applied test for the most fundamental clinical feature of the condition.

A reliance on the visual judgement of human experts for the detection and assessment of Parkinson's is problematic. Human vision cannot make precise measurements of speed and distance [95]. This is likely to limit the accuracy of assessment, particularly for subtle changes present in the earliest stages of disease. In support of this, there is evidence of considerable inter-rater variability for the UPDRS scoring of bradykinesia, e.g. kappa of 0.44 (<0 = no agreement, 1 = perfect agreement) [131]. At a broader level, several studies suggest that human experts also see visual signs of Parkinson's in people without the condition, and when asked to judge video of visual examination signs will frequently misdiagnose tremor conditions, for example confusing Parkinson's with essential tremor or vice versa [91, 92, 96, 132].

In recognition of the limitations of using human expert judgement to assess bradykinesia, the last 25 years have seen a large number of publications reporting the application of technology to record the finger tapping test for bradykinesia. The authors of such studies commonly describe them as "objective" [12-14, 17, 32, 40, 44, 133] and as "quantifying" [17, 21, 26, 31, 32, 43, 134-136] bradykinesia, implying that their technological measurement has overcome the limitations of subjective human assessment. Studies using technology sometimes measure other aspects of examination, such as tremor, and sometimes other bradykinesia examination items, such as upper limb pronation-supination [16, 135]. However, as with clinical practice, so in technology studies, a test of finger tapping is almost always included, and in some cases finger tapping is the only measured test for bradykinesia.

The assumption that technology quantifies bradykinesia may not be correct. Firstly, the initial recording of human movement, or the measurements derived from it, may not always be accurate. For example, accelerometers can show poor accuracy in quantifying human movement [137] and are subject to gravitational artefact [86]. One commercial infrared camera shows poor accuracy below movement of 2cm amplitude [138, 139]. Secondly, measurement in three dimensions is usually summarised into a one-dimensional number (e.g. angular velocity), so some of the original signal is lost. Thirdly, bradykinesia is a concept that derives from expert consensus of what is visible when people with Parkinson's undertake repetitive movements. Human visual judgements and visual perception are inherently imprecise. Thus, it is possible that the actual movement of people with Parkinson's and those without the condition might not correspond very well with the current definition of Parkinson's bradykinesia. If the phenomenon exists to some extent in the minds of expert observers rather than in the movement of those they observe, it becomes difficult to quantify by recording movement. Human perception (including visual perception) is prone to biases and inaccuracies, with schemata and heuristics frequently applied [140] and sometimes 'gestalt' perception [141-143] that cannot be easily expressed as a simple sum-of-parts rule.

This chapter is a scoping review of published technologies that attempt to automate the finger tapping test for bradykinesia in Parkinson's. The topic is an example of attempts to objectively measure an examination sign for a common neurological condition, and so it has a wider relevance beyond Parkinson's. I will describe the different types of technology used to record finger tapping, the measures derived from those recordings, and the main tests applied to those measures, with a focus on the clinical definition of bradykinesia. I will summarise findings relating to measures of tapping speed, amplitude, rhythm and decrement, for group mean differences and correlation with UPDRS (and MBRS). I will also describe attempts to use novel measures that are not simple surrogates for speed, amplitude or rhythm, and the use of machine learning for classification into UPDRS categories or Parkinson's / control.

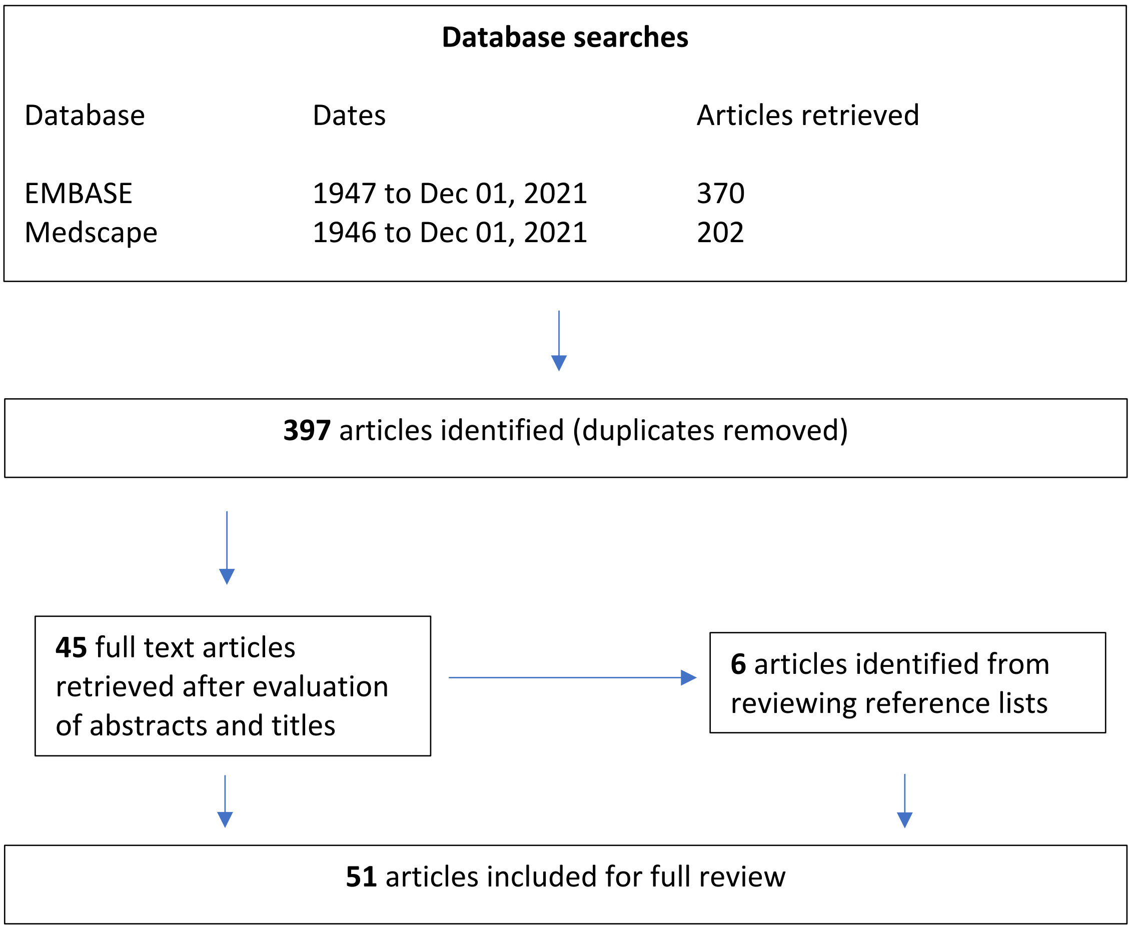
Methods

Box 2.1. Keywords and phrases used in database search

1. Finger-tap* (this includes results for ‘finger tap’ without a hyphen)
2. Parkinson*
3. Bradykine*
4. Akine*
5. Technolog*
6. Quantif*
7. Measure*
8. (#1 AND #2 AND #3) OR (#1 AND #2 AND #4) OR (#1 AND #2 AND #5) OR (#1 AND #2 AND #6) OR (#1 AND #2 AND #7)

*:wildcard indicator
Limited to English language. Excluding reviews and conference abstracts

Box 2.2. Selection of studies for review.



Database searches

Database	Dates	Articles retrieved
EMBASE	1947 to Dec 01, 2021	370
Medscape	1946 to Dec 01, 2021	202

397 articles identified (duplicates removed)

45 full text articles retrieved after evaluation of abstracts and titles

6 articles identified from reviewing reference lists

51 articles included for full review

Search strategy and study selection

I performed a literature search for original research studies, using a combination of keywords (Box 2.1) across a range of databases (Box 2.2), from database inception to August 2021. Only studies published in English were included. Conference abstracts and review articles were excluded.

The titles and abstracts were reviewed from the initial searches. Articles included were those that involved quantification or measurement of finger tapping according to standard instructions ('as fast and big as possible') in people with Parkinson's by technological means, and a comparison of these measurements with either diagnosis category (Parkinson's vs control) or clinician rating of bradykinesia. The testing of finger tapping measurements against diagnosis or clinical rating did not have to be the main focus of the paper for inclusion.

Studies that purely measured usability or compliance were excluded, as were studies that used technology to measure finger tapping purely as an outcome (e.g. of an exercise programme) without any comparison of the results with diagnosis status or clinical rating [144]. Studies using aggregate measures across several clinical tests, without specifying the specific results for finger tapping in isolation were also excluded [99]. Pure comparisons of 'on' and 'off' medication conditions were not included [134]. Two journal articles from 2020 were my own, published forms of the work described Chapter 4 of this thesis [145, 146]. These were excluded from the review, given that they are described fully in Chapter 4.

Data extraction

Data were extracted using the following questions.

1. What is the method used to record finger tap movement ?

2. For each of tapping speed, amplitude and rhythm, what is:

(a) the measurement derived from the recording ?

(a) the comparison of patient and control means ?

(b) the correlation with clinical ratings ?

(c) the classification into patients and controls ?

3. What approaches have been used to combine multiple aspects of tapping into a single measure, and what results have been reported with these ?

Results

A total of 397 articles were found, and after evaluation of abstracts, 45 full text articles were included for review, in addition to 6 further articles identified from abstracts, to make a total of 51 articles reviewed.

Recording the movement: devices to record finger tapping movement

Over the last 25 years, a number of publications have described a range of technological devices to record movement during the finger tapping test. The different approaches include: surface contact, gyroscope/accelerometer (often combined in one device), infrared camera, electromagnetic sensors, and (more recently) computer vision applied to standard video.

Surface contact methods involve participants tapping one or two fingers on a surface (index +/- finger), which produces a binary (on/off) signal based on the presence / absence of the fingertip on the surface over time. Devices have included computer keyboards [12, 13], pressure sensors including musical equipment [14, 147], and more recently smartphone screens [18, 20, 46]. Figure 1 shows two examples of surface contact approaches.

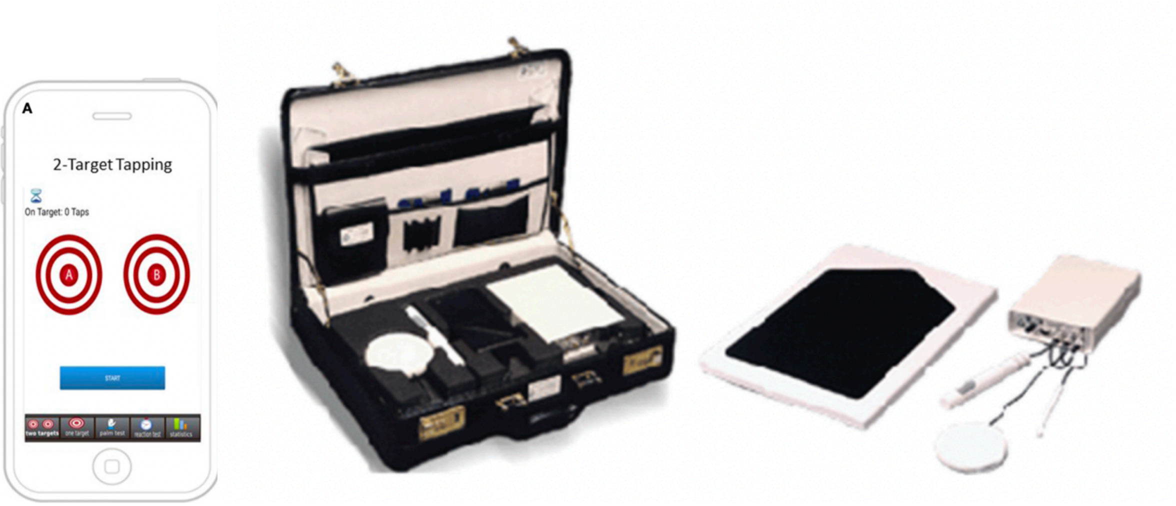


Figure 2.1. Two surface contact methods to record finger tapping. The left side of the figure shows smartphone screen targets [20], the right side shows equipment that includes a touch recording plate (the white circle) [17].

Repetitive surface tapping differs from the standard finger tapping test, in which finger and thumb tip are tapped against each other. It is possible that the absence of thumb involvement might subtly change the resulting performance of the task, and it is known that changes in various external cues (including a fixed target) can alter Parkinson's movement [148]. More importantly, surface contact has only a limited potential to record the basic movement of the clinical finger tapping test. It cannot record the amplitude of the tapping movement, which is one element of bradykinesia. This in turn means that it cannot necessarily record the speed of tapping because velocity of finger movement (distance/time) cannot be determined. Tapping frequency is recorded (number of taps/time), but any given frequency could be due to fast, larger amplitude movements or slower but smaller amplitude movements. Thus, frequency of tapping is limited as a proxy measure of the "speed" of movement.

Gyroscope methods are often combined with accelerometer in one movement-sensing device. A common recording made from such equipment is angular velocity of the finger tip, or finger tip and thumb tip, with the three dimensions of gyroscope movement summed [27].

Infrared devices use two or three cameras that detect infrared light, together with markers attached to the hand (the finger, thumb and sometimes additional locations). The markers are either active, meaning they emit infrared light (connected to a power source), or passive, which means they reflect infrared provided by a separate light source (often integrated into the infrared camera device). Multiple cameras in the device allow determination of the three-dimensional spatial coordinates of the markers. From this, the most common measurement derived is the (Euclidean) finger-to-thumb distance over time [36, 38, 39].

Electromagnetic devices involve an attachment on the finger-tip and thumb-tip. The equipment produces a magnetic field that is detected by the finger-tip sensor, the strength of field depends on the distance from the field source to the sensor. This allows the distance between thumb and finger to be calculated over time [44]

Figure 2.2 shows examples of these three types of 'wearable' device (gyroscope/accelerometer, infrared cameras, electromagnetic sensors). A fundamental limitation of all of them is the requirement for specialised equipment. Clinical assessment of bradykinesia is common and widespread, so a large financial cost would be involved in making these devices part of routine clinical practice (an investment that has yet to occur). They are likely to remain restricted to research settings. Another problem with these devices is that the weight/size or movement restriction involved in attaching equipment to the digits could potentially subtly alter movement.

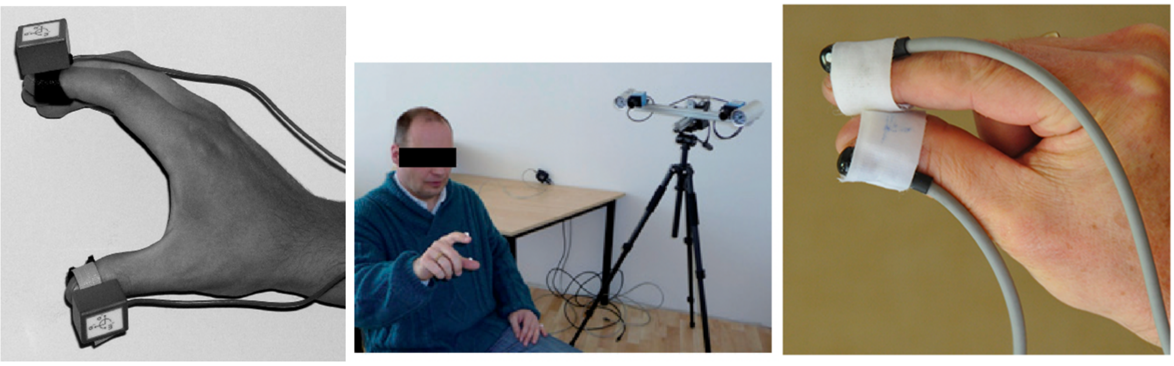


Figure 2.2. Wearable devices to record finger tapping. Left to right: systems based on gyroscopes [27], infrared cameras [37], and electromagnetic sensors [44].

In the period covered by the literature search, two research groups have used computer processing of standard video to record finger tapping movement. This is a form of computer vision. It requires no special equipment, because it uses only a standard video camera. It is based on computing algorithms that can track moving objects in video. In one publication, a region of the video frame containing the tapping hand is first isolated by using previously developed algorithms that detect the human face [149]. The participants positioned their tapping hands either side of their face, such that face detection enabled isolation of the adjacent tapping hand regions, and the amplitude of tapping could be normalised to face size. Algorithms that detect pixel movement magnitude and direction were then applied to that isolated region, to derive a measure of relative finger-tip movement [149]. The other published video report uses the computer vision technique 'pose estimation' [135]. In this, as outlined in Chapter 1, algorithms are applied that recognise and track the position of body parts across video frames (in this case hand, fingers and thumb). The recognition is based on applying previously developed convolutional neural network algorithms that have been trained on large datasets of images.

Although video does not require special equipment, there are several potential disadvantages as a method to record finger tapping. Firstly, it only records movement in two dimensions, so that movement involving a component perpendicular to the camera view will be underestimated. Secondly, there is no absolute measure of distance. This means that amplitude and velocity have to be calculated in relative terms, as a proportion of maximum amplitude [135]. Alternatively, a distance marker needs to be taped to the wall and filmed in the video frame (introducing an element of equipment and calibration that could limit widespread use in clinical settings) [150]. However, in practice, relative amplitude may be enough information to recognise movement abnormalities (which are specified in relative terms in international criteria, not in absolute cm or cm/s). Thirdly, in order to track finger tapping movement, the methods require manual (human) labelling of the points to be tracked (e.g. finger-tip, thumb-tip, and other landmarks) in a series of video frames from the set to be analysed. Currently, this involves manual labelling of a minimum of hundreds of video frames, and does not generalise to a new set of videos (the labelling has to be done for each set of videos to be analysed). Thus, at present it is highly labour-intensive. In the future, this step is could potentially become fully automated, though this has not yet been achieved.

The signal processing: measures derived from finger tapping recordings

Many of the devices employed to record finger tapping capture movement in multiple dimensions, and often movement of more than one part of the hand. For example, gyroscope devices usually record movement in six degrees of freedom, and infrared cameras can measure movement in three dimensions, often with markers on multiple locations on the hand. However, all of the finger tapping studies published to date involve an initial step that reduces the recording to a one-dimensional signal or a small set of one-dimensional signals (e.g. [32, 37, 40]). In other words, they begin by reducing the recording to a time-series, in which the x-axis is time and the y-axis is a single variable, such as finger-to-thumb distance. As such, much of the recorded movement information is discarded at the outset.

The basic time-series is often similar in different studies. The initial measure in surface tapping studies is whether the finger-tip is on or off the surface at each time point. Gyroscope studies start with finger-tip angular velocity over time, either in one plane or in three planes summed together [151]. Electromagnetic studies measure finger-tip to thumb-tip distance over time [44]. Video studies have also measured finger-tip to thumb-tip distance over time but as a proportion of maximum distance (i.e. relative distance) [135].

From these one-dimensional time series, researchers then derive a number of features that they judge to reflect the elements of bradykinesia. Certain features are commonly used across many studies. In surface tapping studies tapping frequency (taps per unit of time) is used as a measure of speed [19]. With other devices, velocity of finger or finger-to-thumb movement is the speed measure [47]. To measure tapping amplitude, it is excursion angle using gyroscopes [27] and finger to thumb distance with electromagnetic and video studies [149].

Across all devices, rhythm is often measured by calculating standard deviation or coefficient of variation (standard deviation divided by the mean) of the mean tap interval (e.g. [39]). More unusual measures of rhythm have also been derived. An accelerometer and gyroscope study derived a ‘smoothness index’ of tapping, from the spectral arch length of the movement speed profile [34]. The MDS-UPDRS specifies impairment in rhythm as “interruptions” and “hesitations” [8]. In contrast to other studies, Bank et al devised measures of rhythm impairment in infrared recordings that were closer to that UPDRS wording [40]. They used a peak detection algorithm to identify individual tapping cycles, and defined 'hesitations' as the number of zero velocity crossings minus the 2 reversals per cycle. 'Halts' were defined as periods longer than the average duration of 2 cycles in which all data points were within 1.5cm. In one of the earliest video studies, Khan *et al* derived a measure of tapping rhythm termed 'cross-correlation of the normalised peaks' [149]. This involved splitting the recording into two time slots, 2s to 5s and 6s to 9s, then constructing a 'peak signal' by joining the tapping peaks in each time slot (i.e. a line joining neighbouring peaks). Cross-correlation was then applied to measure the similarity between the peak signals of the two time slots (i.e. how regular the tapping rhythm was) [149].

The MDS definition of bradykinesia refers to "decrement in amplitude or speed" [7]. Thus, the current concept of bradykinesia specifies progressively smaller movements or progressively slower movements, as movement continues. This is distinct from movement that is slow or small from the outset but remains stable over the period of tapping. Similarly, the MDS-UPDRS finger tapping subscore wording also specifies abnormal amplitude with the words "amplitude decrement", and the earlier in the 10 taps that the amplitude decrement occurs, the higher the MDS-UPDRS grade [8]. In this context, it could be seen as inappropriate that the measures described so far reflect overall mean or maximum speed and amplitude, rather than measures of decrement. However, a minority of studies have, perhaps more appropriately, attempted to derive a decrement measure from finger tapping recordings. The most common method to look for decrement is to calculate the slope of a linear regression line fitted to the time series of tapping speed or tapping amplitude [40]. Other methods compare early and late sections of the tapping recording [35, 45].

In more recent studies, tapping features are combined into a new composite measure, that is not a simple direct measure of one aspect of the tapping recording. This approach usually involves the application of machine learning techniques [28, 42, 43].

Group mean differences – speed, amplitude and rhythm

A number of technology studies have taken variables derived from finger tapping and compared the group means of those variables for people with Parkinson's (PwP) versus controls.

There have been 10 studies that reported group mean tapping frequency (number of taps / time) to be lower in PwP compared with controls (7 surface contact, 1 gyroscope, 2 infrared camera) [12-14, 17, 20, 21, 35, 40, 46, 151]. However, 5 studies found no significant group difference, i.e. p>0.05 (2 surface contact, 1 gyroscope, 2 infrared camera) [18, 23, 38, 47]. In a strain-gauge goniometer experiment and one gyroscope study, there was only a group difference when recordings were made 'off' participants usual Parkinson's medication [34, 45]. The mean tapping frequency of participant groups in different studies varies widely, and comparing across studies control group means can be slower than PwP group means, suggesting there is no neat cut off for what can be considered an 'abnormal' tapping frequency, Figure 2.3. This is not easily explained by protocol differences in the total duration of tapping, Figure 2.4, but likely reflects tapping frequency as a poor measure of speed, because tapping frequency cannot distinguish high frequency secondary to small amplitude taps from that secondary to high velocity of finger movement.

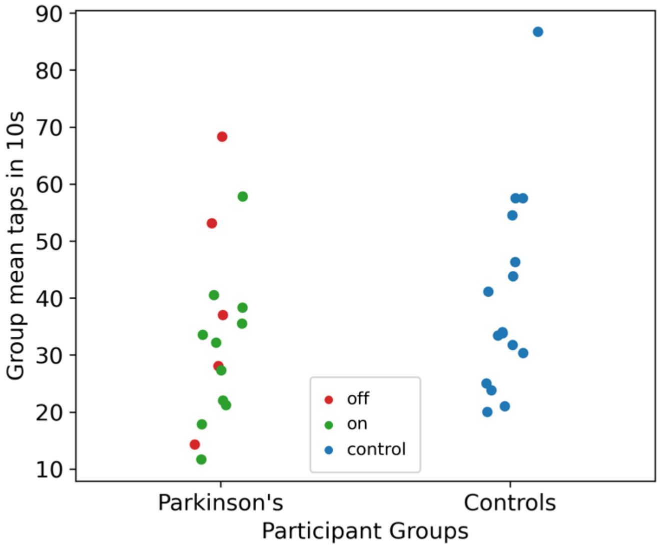


Figure 2.3. The group means for tapping frequency vary widely across studies, for both Parkinson's group means and control group means, and many control means are slower than Parkinson's means. Dots show published group means for tapping frequency (standardised to taps/10s) in PwP on usual

medication, ‘on’ (green), with medication withheld, ‘off’ (red), and controls (blue) [12-14, 17, 20, 21, 25, 34, 35, 38, 40, 45-47, 151].

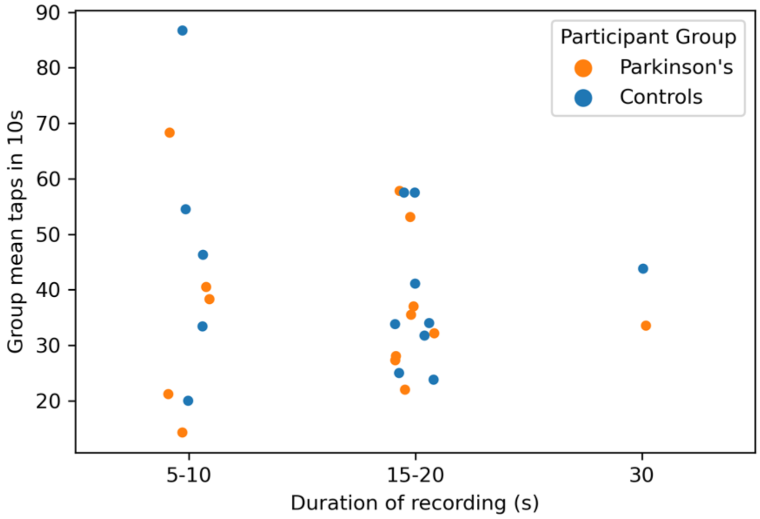


Figure 2.4. Variation in group mean tapping frequency across studies is not explained by variation in protocols for total tapping duration. Dots show published group means for tapping frequency (standardised to taps/10s) in PwP (orange) and controls (blue), according to the total duration of tapping recorded in the study protocol. Note PwP means are often faster than controls and vice versa, and this holds across different total tapping durations [14, 17, 20, 21, 25, 34, 35, 38, 40, 45-47, 151].

Group mean tapping velocity (distance / time) was slower in PwP compared with controls in 1 gyroscope study and 4 infrared camera studies [38-40, 47, 151], although recordings were made ‘off’ Parkinson’s medication in one of the infrared protocols [47]. It was calculated as either the overall velocity or the maximum opening velocity.

Group mean tapping amplitude (degrees or cm) was significantly lower in PwP compared with controls in 1 gyroscope study and 2 infrared camera studies [35, 40, 151]. No significant difference was found in a third infrared study [47].

Impairment of finger tapping rhythm, measured as the standard deviation of either tap interval or tap rate, was higher (more irregular rhythm) at the group mean level for PwP compared with controls in 3 surface contact studies, but not in a fourth [12, 13, 17, 152]. The coefficient of variation of tapping interval (standard deviation divided by the mean) showed a higher group mean (more irregular rhythm) in PwP compared with controls in one accelerometer study and one gyroscope study [25, 151]. An infrared study only found such a difference from controls in a participant group with advanced stage Parkinson's but not in one with early-stage Parkinson's [39]. Coefficient of variation can also be applied to tapping velocity. It was higher in PwP in one gyroscope study and one infrared study, but the latter used recordings made with participants 'off' their usual Parkinson's medication [47, 151].

Spectral arch length of movement speed profile (a rhythm measure) was significantly different between controls and PwP groups, using gyroscope/accelerometer, but only when Parkinson's medication was with-held ('off' condition) [34]. An infrared study found that group mean tapping hesitations were higher in the PwP group (0.8 per cycle) vs the control group (0.4 per cycle) [40]. They also reported both controls and PwP to have a mean 'halt' value of 0, though the interquartile range was 0-0 in controls, and 0-21.6 in PwP [40]. That suggests the data was not normally distributed, so the mean would be an inappropriate measure.

Correlation with UPDRS – speed, amplitude and rhythm

Many studies that have used technology to record finger tapping have tested correlation between the measurements they obtain and clinician ratings according to the Unified Parkinson's Disease Rating Scale (UPDRS). Two parts of the UPDRS are most commonly tested. One is finger tapping item (3.4), UPDRS-FT, which is scored from 0-4, **Table 1.1**. The other is the motor examination ('Part 3') subsection score, UPDRS-ME, which is the total of 18 motor examination items, each with a 0-4 score. There is a wide variation in the reported correlation strengths, **Table 2.1**.

Authors, year, participants	Method	Element of tapping	Clinical rating	Correlation	Caveats / notes
Giovanonni et al 1999	s	Frequency	UPDRS-ME	r -0.69	Recorded 'off' medication
		Rhythm	UPDRS-ME	r 0.41	
Homann et al 2000	s	Frequency	UPDRS-ME	r -0.60	Recorded 'off' medication
Tavares et al 2005	s	Frequency	UPDRS-ME	r 0.50	Recorded 'off' medication
		Frequency	UPDRS-B	r 0.67	
		Rhythm	UPDRS-ME	r 0.56	
Papapetropoulos et al 2010	s	Frequency	UPDRS-FT	r NS	
		Frequency	UPDRS-B	r NS	
Maetzler et al 2015	s	Frequency	UPDRS-ME	r NS	More than half the patients recorded 'off' medication
		Frequency	UPDRS-FT	r NS	
		Rhythm	UPDRS-FT	r 0.13	
		Rhythm	UPDRS-ME	r 0.15	
Lee CY et al 2016	s	Frequency	UPDRS-ME	r -0.37	
		Frequency	UPDRS-B	r -0.75	
Kassavetis et al 2016	s	Frequency	UPDRS-FT	r -0.75	All patients 'off' medication
Mitsi et al 2017	s	Frequency	UPDRS-ME	r NS	
Roalf *et al.* 2018	s	Frequency	UPDRS-ME	r NS	
Yokoe *et al.* 2009	a	Velocity	UPDRS-FT	r -0.59	Controls assumed to be UPDRS-FT grade 0
Costa *et al.* 2010	a	Frequency	UPDRS-FT	r NS	First 2 seconds of tapping omitted
		Amplitude	UPDRS-FT	r NS	
Kim *et al.* 2011	g	Frequency	UPDRS-FT	r -0.72, -0.74	Some control means worse (slower) than patients

Authors, year, participants	Method	Element of tapping	Clinical rating	Correlation	Caveats / notes
		Velocity	UPDRS-FT	r -0.66, r -0.70	
Heldman *et al.* 2011	g	Velocity	MBRS Speed	r -0.79	Half recordings 'off' medication Inappropriate correlation method (Pearson)
		Amplitude	MBRS Amplitude	r -0.81	
		Rhythm	MBRS Rhythm	r 0.65	
Lee *et al.* 2015	g	Frequency	UPDRS-FT	r2 0.74	Early stage disease
		Velocity decrement	UPDRS-FT	r2 NS	
		Rhythm	UPDRS-FT	r2 0.47	
Ling *et al.* 2012	i	Velocity	UPDRS-ME	r -0.68	Half recordings 'off' medication. Amplitude worse (smaller) in healthy controls than PD off medication.
		Velocity decrement	UPDRS-ME	r NS	
		Amplitude	UPDRS-ME	r -0.79	
		Amplitude decrement	UPDRS-ME	r NS	
		Rhythm	UPDRS-ME	r 0.75	
Ruzicka *et al.* 2016	i	Frequency	UPDRS-FT	r NS	2 rater cohen k = 0.59 UPDRS-FT median HC 1.0 and PD 1.5
		Velocity	UPDRS-FT	r 0.48	
		Amplitude decrement	UPDRS-FT	r NS	
Bologna *et al.* 2016	I	Velocity	UPDRS-ME	r NS	Late stage PD subgroup were recorded 'off' medication Only 1 clinical rater
		Velocity decrement	UPDRS-ME	r NS	
		Amplitude	UPDRS-ME	r NS	
		Amplitude decrement	UPDRS-ME	r NS	
		Rhythm	UPDRS-ME	r NS	
Bank *et al.* 2017	I	Frequency	UPDRS-FT	r -0.24	Low test-retest reliability for speed
		Frequency decrement	UPDRS-FT	r NS	
		Velocity	UPDRS-FT	r -0.37	

Authors, year, participants	Method	Element of tapping	Clinical rating	Correlation	Caveats / notes
		Velocity decrement	UPDRS-FT	r NS	
		Amplitude	UPDRS-FT	r NS	
		Amplitude decrement	UPDRS-FT	r NS	
		Rhythm	UPDRS-FT	r 0.31, 0.27	
Teo *et al.* 2013	o	Frequency decrement	UPDRS	r 0.81	Half PD recordings 'off' medication
		Amplitude decrement	UPDRS	r 0.76	
Khan *et al.* 2014	v	Frequency	UPDRS-FT	Guttman -0.54, -0.38	Grade 4 UPDRS-FT excluded. PD group older than controls (50-75 vs 40-60)
		Frequency decrement	UPDRS-FT	Guttman 0.62, 0.44	
		Velocity	UPDRS-FT	Guttman -0.59, -0.38	
		Amplitude decrement	UPDRS-FT	Guttman -0.84, -0.41	
		Rhythm	UPDRS-FT	Guttman -0.80, -0.40	
Liu *et al.* 2019	v	Frequency	UPDRS-FT	r 0.912	Exceptionally strong correlations despite measures very similar to other studies Pearson correlation inappropriate for ordinal data
		Amplitude	UPDRS-FT	r -0.935	
		Rhythm	UPDRS-FT	r 0.815	
		Amplitude variability	UPDRS-FT	r -0.388	

Table 2.1. Published correlations between technology measures of finger tapping bradykinesia elements and clinical ratings. Note the very wide range of correlation strengths. Method column, s: surface tapping, g: gyroscope, i: infrared, v: video. r = correlation coefficient. UPDRS-FT: Unified Parkinson's Disease Rating Scale Finger Tapping Subscore. MBRS: Modified Bradykinesia Rating Scale. [12, 13, 15, 17-20, 23-27, 38-40, 45-47, 135, 149, 151]

There is an enormous variation in the reported strengths of correlation between measures of tapping frequency and UPDRS ratings. Four studies found no significant correlation between tapping frequency and UPDRS-FT scores (two surface contact, one accelerometer, one infrared) [17, 18, 25, 38]. Another five studies found the following statistically significant but wide-ranging correlation strengths between tapping frequency measures and UPDRS-FT: -0.16 (video), -0.24 (infrared), -0.34 (gyroscope), -0.38 (video), -0.74 (gyroscope), -0.75 (surface contact), -0.91 (video) [19, 34, 40, 149, 153]. There are similarly varied findings for the correlation between tap frequency and UPDRS-ME. Three surface tapping studies reported no significant correlation [18, 20, 23], while another four surface tapping studies reported significant correlations ranging in strength from -0.37 to -0.69 [12, 13, 46, 147].

In terms of measures of tapping velocity (distance over time), the following correlations with UPDRS-FT ratings have been reported: -0.37 (infrared), -0.38 (video), -0.48 (infrared), -0.59 (accelerometer), -0.59 (video), -0.68 (infrared), -0.72 (gyroscope), -0.74 (gyroscope) [24, 26, 38, 40, 47, 149]. Only one study has tested UPDRS-ME correlation with tapping velocity, and found no significant correlation [39].

Mean tapping amplitude showed no significant correlation with UPDRS-FT in one infrared study [40], but in three other studies there were significant correlations: -0.66 (gyroscope), -0.70 (gyroscope), -0.935 (video) [26, 153]. On testing correlation with UPDRS-ME, one infrared study found no correlation [39] but another reported -0.79 correlation [47].

Measures of tapping rhythm have been tested for correlation with UPDRS items. Standard deviation of intertap interval has been reported to correlate with UPDRS-FT 0.13 in a surface contact study, and 0.82 in a video study [18, 153]. Standard deviation and UPDRS-ME showed no significant correlation in one study and 0.15 correlation in another (both surface contact) [17, 18]. Another surface contact study reports “variance of the time interval between key strokes”, which is presumably standard deviation, to correlate 0.41 with UPDRS-ME [12]. Coefficient of variation (standard deviation divided by the mean) of intertap interval showed no significant correlation with UPDRS-FT in an accelerometer study [25], and no significant correlation with UPDRS-ME in an infrared study [39]. However, in a surface contact study, coefficient of variation of tap interval correlated 0.56 with UPDRS-ME. Coefficient of variation of key strike duration has been reported to correlate 0.67 with the three UPDRS bradykinesia items [147]. Coefficient of variation of tapping velocity (rather than intertap interval) correlated 0.75 with UPDRS-ME in a report using infrared camera [47].

In a gyroscope/accelerometer study, the spectral arch length of movement speed profile (a rhythm measure) correlated significantly with UPDRS, with r2 of 0.47 (r is not given) [34]. In a video study that used the cross-correlation of the normalised peaks as a measure of rhythm, the average (across several tapping trials) showed a Guttman correlation of 0.80 with UPDRS-FT for one rater, but only 0.40 for a second rater [149]. In infrared recordings the number of tapping hesitations correlated 0.31 with UPDRS-FT, while the percentage of time in 'halts' correlated 0.27 with UPDRS-FT [40].

Very few studies have compared finger tapping recordings with the modified bradykinesia rating scale (MBRS), a scale that asks the human rater to separately grade tapping speed, amplitude and rhythm, producing three separate scores, each between 0 and 4, Table 1.2. Heldman *et al* used a gyroscope method to measure tapping angular velocity, excursion angle and coefficient of variation, reporting correlations of -0.79, -0.81 and 0.65 with MBRS speed, amplitude and rhythm ratings respectively [27]. However, they converted ordinal ratings into continuous numbers (e.g. "MBRS speed 2.7") and calculated means and Pearson correlations, all inappropriate methods for analysing ordinal data. Half their recordings were in the 'off' medication state.

Measures of decrement

Fitting a linear regression line to tapping frequency has been used as a measure of speed decrement, by comparing the slope of the line in participant groups. However, by this method, neither gyroscope nor infrared methods found a significant difference between PwP and control groups in the slope of the line. For a regression line to measure velocity decrement, one gyroscope study found a significantly steeper, negative line (suggesting greater decrement) in PwP compared with controls [151]. However, one gyroscope study and three infrared studies found no group mean difference in the slope of the regression line for velocity, comparing PwP with controls [34, 39, 47]. For tap amplitude, two studies (gyroscope and infrared) found a greater (more negative) mean slope of the linear fitted line in PwP compared with controls [47, 151], but two infrared studies did not [36, 47]. A fifth study, using infrared, reported that the “decrease in maximal opening distance” was significantly greater in PwP than in controls, but the details of this measure are not described [38]. Finally, another infrared study found the group mean slope of the regression line was significantly more negative than controls in early-stage Parkinson’s, but not in advanced Parkinson’s [39].

An infrared study used a different method, comparing the duration and amplitude of the final three taps with that of the first three taps. The authors found a significantly longer duration and significantly smaller amplitude of the final three taps in PwP but not in controls [35].

Researchers have also tested for correlation between decrement measures and UPDRS ratings. In most studies, the slope of a regression line is used. For tapping velocity, one infrared method reported no significant correlation between regression line slope and UPDRS-ME [47]. A report using gyroscopes reported that velocity decrement correlated 0.49 with UPDRS, but only when tested with usual Parkinson’s medication withheld. Tested ‘on’ medication, no significant correlation was found [34]. For regression lines testing amplitude decrement, two infrared studies reported no correlation with UPDRS-ME [39, 47] and one infrared study found no correlation with UPDRS-FT [40]. A different approach to amplitude decrement has been to measure the decrease from maximum opening distance in the initial 5 taps, using infrared recordings. No correlation was found between this measure and UPDRS-ME [38]. In stark contrast to the lack of correlation in other publications, Teo *et al* reported that amplitude decrement measured by strain gauge goniometer correlated 0.74 with UPDRS scores, though they used Pearson correlation, which is inappropriate for ordinal (UPDRS) data. Their method divided 20 seconds of tapping into 2 second epochs and then compared amplitude in the first and last epochs [45]. Recordings were made with 11 participants, all ‘off’ their usual Parkinson’s medication.

Novel single measures

Several studies have derived other measurements from recordings of finger tapping, which are not simple surrogates for tapping speed, amplitude, decrement, or rhythm.

Prince *et al* undertook a large study of smartphone screen tapping, with 312 PwP and 236 healthy controls, who each completed 20 separate repeated trials of tapping. They found that PwP showed more longitudinal variability and a greater number of test instances to reach steady state [22]. As with other measures, these are group mean differences, not absolute differences. In another surface contact study, Roalf *et al* found that the control group showed a lower mean intra-individual variability across trials, compared with PWP [23]. Costa et al's accelerometer study derived a new measure from finger tapping: 'beat decay of the auto mutual information value'. This estimates signal predictability by measuring loss of signal over a timescale. Mean values were significantly higher in PwP compared with controls, and the measure showed a correlation of 0.45 with UPDRS-FT scores [25]. They found even higher mean values for essential tremor patients (i.e. PwP mean scores were between those for controls and essential tremor), suggesting the measure by itself is not specific to Parkinson's bradykinesia.

Two studies derived power measurements from a tapping velocity signal, after a Fourier transformation to convert the signal's time domain to the frequency domain. This can be seen as measuring the intensity of movement. Kim *et al*'s gyroscope study found that peak power correlated -0.80 with UPDRS-FT, while total power correlated -0.78 [26]. However, Di Biase *et al* also used a gyroscope and found no significant difference in total power between controls and PwP taking their usual medication, only finding a significant difference when PwP were 'off' usual medication. They quote an R2 correlation of 0.57 with UPDRS-FT scores, but this is PwP 'off' medication [34].

Logistic regression and area under the receiver operating characteristic curve

Another approach to the data recorded from finger tapping is to perform a logistic regression: a logistic function to model a binary dependent variable, in this case whether the tapping is performed by a PwP or control. The result can be represented as a receiver operating characteristic (ROC) curve. This is a plot of the true positive rate against the false positive rate. The area under the curve (AUC) for the ROC is the probability that the logistic regression model will rank a randomly chosen positive instance (PwP) higher than a randomly chosen negative instance (control). The maximum value for AUC is 1 (or 100%), which would indicate a perfect test for PwP vs control in that group of participants (100% sensitive and 100% specific). An AUC of 0.5 (50%) would suggest that the model for classification as PwP or control performs no better than chance. Several studies have reported AUC for measures taken from finger tapping recordings, and also for models that combine multiple measures of different aspects of the tapping signal.

Mitsi *et al.*'s smartphone screen tapping study performed logistic regression on each of a variety of tapping measurements. Their highest AUC figure was 90% for reaction time. They reported an AUC of 83% for two-target interval, a measure of tap frequency [20]. The same study also reported a combined model of several tapping features (two-target total taps, reaction time, two-target total accuracy) that produced maximum discriminatory performance between PwP and controls, with an overall AUC of 0.98 (sensitivity 94%, specificity 93% [20]. In another smartphone study, tapping frequency showed AUC of 0.59 for PwP/control discrimination, while intra-individual variability across trials showed AUC of of 0.89 [23]. Using an accelerometer, the measures of frequency produced an AUC of 66.9 % for PwP/control [25].

Using infrared cameras, Krupicka *et al*. reported discrimination of PwP from controls with AUCs of 0.77 for opening velocity (speed) and 0.87 for decrease in maximal opening distance (amplitude decrement). When combining these two measures into a multivariate logistic regression they found AUC 0.94 [37]. Half the recordings were made with PwP 'off' medication. In a second study from the same group, the same two measures (opening velocity and decrease in maximum opening distance) showed similar AUCs of 0.81 and 0.87 [38].

Yokoe *et al.* performed logistic regression for tapping variables measured by accelerometer, but reported Akaike Information Criterion and misclassification rate, rather than AUC. They report misclassification rates of 15.6%, 18.8% and 32.3% for measures of speed (maximum opening velocity), amplitude (total distance of finger movement) and rhythm (standard deviation tap interval) respectively [24].

Machine learning for classification

Machine learning can be defined as "a subfield of computer science that is concerned with building algorithms which, to be useful, rely on a collection of examples of some phenomenon" [154]. Several studies have applied machine learning techniques to measurements derived from recording finger tapping, most commonly with the aim of a method that will discriminate PwP from controls, or discriminate different levels of bradykinesia.

Yokoe *et al.* derived 14 parameters of finger tapping in 16 PwP and 32 controls by recording with an accelerometer and touch sensor. They plotted these as radar charts, showing different patterns of mean values. They also applied principal component analysis to the 14 parameters. As described in Chapter 1, this is a technique that reduces the dimensionality of the data, aiming to derive new meaningful variables for FT performance (the principal components). They found that the first three principal components together accounted for 80% of the variance of the original data. The first component was largely composed of mean and standard deviation of amplitude and velocity, the second component number of taps and inter-tap interval, and the third component standard deviation of finger tapping intervals [24]. It is unclear what practical implication these results have. The top three principal components involve speed, amplitude and rhythm, as would be expected, and they did not report a logistic regression result for the principal components.

Stamatakis *et al.* used an accelerometer to record finger tapping in 36 PwP and 10 control participants. They epoched the raw signal to isolate successive finger tapping movements, and then selected 18 movement features from the resultant data (e.g. mean opening angle, mean maximum opening acceleration etc). A 'greedy backward selection algorithm' was used with ordinal logistic regression to select the most relevant features in the prediction of MDS-UPDRS FT scores made by 3 clinical raters [28]. They report a Goodman-Kruskal Gamma Index of 0.961 between the UPDRS-FT scores predicted by the logistic regression model and the mean rater scores (rounded to the nearest integer). Goodman-Kruskal is a test of rank correlation. They obtained similarly very high Goodman-Kruskal gamma index scores for the rank correlations between the human clinical raters (0.87 to 0.97), despite previously published measures of interrater agreement for UPDRS-FT suggesting only moderate to good agreement (e.g. intraclass correlation of 0.58 [32]). This perhaps suggests a tendency for Goodman-Kruskal to produce higher results *per se* with UPDRS-FT. They also reported AUC for binary UPDRS-FT classifications as follows: 0 vs 123 of 0.945, 01 vs 23 of 0.919, 012 vs 3 of 0.970. The authors excluded the six UPDRS-FT grade 4 scores in their study, "as by definition they were not able to perform the task”, meaning that the classification task was only for 4 categories, not 5 as per standard UPDRS-FT.

Kim *et al.* derived 14 variables from finger tapping measured by gyroscope in 40 PwP. They performed stepwise multiple linear regression using these tapping features for each of two clinical raters. The ability of each regression model to discriminate different clinical scores was evaluated by ANOVA and post-hoc analysis. In the post-hoc comparisons, regression models could discriminate all pairs of UPDRS-FT scores, with one exception. In contrast, single tapping features (e.g. coefficient of variation of angular velocity) could not distinguish most neighbouring pairs of UPDRS-FT scores, and many non-neighbouring pairs [31].

Martinez-Manzanera *et al.* took 21 features of finger tapping measured with a 9 degrees of freedom sensor (accelerometer, gyroscope and magnetometer) and performed backward linear regression. This obtained a model for performance using averaged UPDRS-FT scores from four clinical raters, although the method inappropriately converted ordinal rating data to continuous data. Leave-one-out cross validation was employed to estimate the model’s performance. The model accounted for 79% of the total variance in the data. Their model showed a cross validation of “0.38” (the units being UPDRS-FT) which was “0.19 less than the average difference in scores across all pairs of investigators” [32]. UPDRS-FT is an ordinal scale, not an interval scale, and thus splitting it into fractions such as 0.39 and 0.19 is difficult to justify and interpret. For example, the difference between UPDRS 0 and 1 is not necessarily the same size as that between UPDRS 2 and 3, the severity of bradykinesia does not necessarily progress smoothly and continuously between each UPDRS grade.

Lainscsek *et al.* recorded finger tapping with an infrared camera system in 13 PwP ‘off’ their usual medication and 13 controls. They used a genetic algorithm to select a nonlinear delay differential equation and then fitted this to each time series so that its coefficients were used as a six-dimensional numerical descriptor. Differential delay equation (DDE) scores showed a 95% confidence interval of 1.05 when plotted as Bland-Altman comparison against mean UPDRS. Again, this is an inappropriate use of mean as a measure for ordinal UPDRS data. The DDE scores correlated 0.785 with ‘mean UPDRS-FT’ [36].

Sano *et al.* recorded finger tapping with an electromagnetic system in 31 PwP and 360 healthy controls. They extracted 21 characteristics from the resultant waveforms and normalised these by age, using regression lines. Principal component analysis was applied to the waveform characteristics in the 31 PwP. The first two principal components reflected motion wideness (amplitude) and variation of wideness and frequency (rhythm). Multiple linear regression with stepwise variable selection was undertaken with the principal component scores and UPDRS-FT, to produce an equation for calculating a finger tapping severity score. Using the leave one out method, this score showed a mean square error of "0.45" UPDRS-FT – the mean value of squared errors between the score and UPDRS-FT [43].

Gao *et al.* applied an evolutionary algorithm to finger tapping data collected, using electromagnetic sensors, from 107 PwP, 47 people with essential tremor, and 49 control participants, to produce a score from -1 to +1 for finger tapping bradykinesia. They report a correlation with UPDRS-FT of 0.819 for the right hand and 0.783 for the left hand. For the classification of tapping as PwP or control, the algorithm gave an AUC of 0.899 [44].

Khan *et al.* derived 11 tapping features from video tracking of finger tapping. 13 PwP and 6 controls performed repeated trials over several days, to make a total of 387 PwP videos and 84 control videos. They used a support vector machine technique to create decision boundaries between three levels of UPDRS-FT: 0, 1 and a merged category of 2 and 3 (there was no grade 4). Ten-fold cross validation showed a classification accuracy of 88% (quoted in the paper's abstract) for one rater, and 76% (not quoted in the abstract) for a second rater, Using the same support vector machine decision boundary technique, they reported an accuracy of 95.8% for classification between PwP and control tapping videos, although there were 14 false positive samples in the (84 video) control group [149].

In another video study, Liu *et al.* applied a support vector machine decision boundary technique to four features of finger tapping: mean tap interval, mean tap amplitude, and their respective standard deviations. This involved 60 PwP (no controls), 360 video clips and 2 clinical raters. They report 'precision accuracy' (analogous to specificity) of 73% for UPDRS-FT grade, and 'recall accuracy' (analogous to sensitivity) of 95% [135].

Discussion

In summary, a number of methods have been used to record movement during the finger tapping test in people with Parkinson's (PwP), including surface contact devices, gyroscope/accelerometer, electromagnetic devices, infrared cameras and standard video. Across all of these, there is an initial step of signal processing in which three-dimensional movement is reduced to a one-dimensional time series, for example, finger to thumb distance (y axis) over time (x axis). From the resultant one-dimensional time series, a number of further calculations are performed, to produce measures that usually have a theoretical basis in the current definition of bradykinesia and its subcomponents (tapping speed, amplitude, decrement, rhythm). These measures are then used to test for group mean differences, correlation with clinician rating by UPDRS, classification by logistic regression or classification by more complex machine learning techniques.

It could be argued, as their authors do, that these studies demonstrate "objective measurement" or "quantification" of bradykinesia, at least to some extent. Multiple papers report that tapping measures show statistically significant mean differences between PwP and controls, as well as significant correlations with UPDRS (sometimes of moderate or high strength, though interpretation of correlation strength involves inherently arbitrary thresholds). Furthermore, classification of PwP/control, or classification into UPDRS score bands, by logistic regression or more complicated machine learning techniques, has produced some accuracy results that are high or appear high enough to be clinically useful. However, all of the studies that attempt to use technology to measure finger tap bradykinesia have fundamental limitations, that bring into question the idea that objective quantification has been demonstrated.

Group mean differences in finger tapping measures do not demonstrate that a technology quantifies bradykinesia. In all measures of finger tapping, the PwP and control group results overlap considerably. There is no absolute difference between normal and pathological. Some controls can show slower, smaller or more irregular tapping than some PwP. An analogy is group mean height for men versus women. The group mean height for men is significantly higher than that for women, but height is not a good measure of whether an individual is male or female, and nor does height quantify the phenomenon of being male. Rather than the mean differences, perhaps the most important result from group comparison is that performance on specific finger tapping measures always overlaps considerably between PwP and controls, suggesting that individual elements of bradykinesia are strikingly non-specific.

For correlation of technology measures with UPDRS ratings, it is difficult to draw conclusions. A common justification for technology to measure finger tapping is that human judgement is subjective, and yet a common method to demonstrate objective technology measurement is to test how well it correlates with (subjective) human ratings. It is hard to understand an approach that attempts to demonstrate objective measurement by its strength of correlation with subjective measurement. Several studies suggest considerable inter-rater variation in human ratings of bradykinesia (a topic investigated in Chapter 3). Having said this, a different approach does support a degree of objective quantification with technology. Adjusting the strength of deep brain stimulation provides a way to improve Parkinson's movement impairment in a stepwise fashion. Heldman *et al.* reported higher test-retest reliability (intraclass correlation coefficient) and sensitivity (minimal detectable change) with a gyroscope measure versus clinician ratings [29].

Correlation studies have additional limitations. Most compare technology measures against clinical ratings made by a small number of people, usually one, two or three raters. It is unlikely that such a low number of raters will adequately capture the median of the range of scores across raters in general. In addition, many studies only feature PwP, and not control participants. This would be expected to remove subtle judgements at the lower end of the scale, between normal and mild abnormality, presumably artificially strengthening correlation. In addition, many studies take place with Parkinson's medication with-held at the time of recording ('off' state). This would be expected to artificially exaggerate bradykinesia, likely also exaggerating abnormalities and making correlation stronger. Confidence in any conclusions is undermined further by widespread use of incorrect statistical methods. It is common for the ordinal of UPDRS-FT (or MBRS) to be converted into a continuous scale (e.g. scores of "2.6"), followed by calculation of mean values and Pearson correlation coefficient, all of which is fundamentally incorrect for ordinal data.

Related to the points above is one striking feature of the results taken as a whole: there is a wide variability of the results across studies. For any given mean difference or correlation, some studies show statistically significant results, others do not, some show large mean differences, others small mean differences, some studies show strong corelations, others show much weaker correlations. Although attempts could be made to partly explain this variation by differences in participants and protocols, it seems surprising that detection of elements of bradykinesia should be such a fragile process. If these are studies showing objective measurement of the core motor feature of Parkinson's, then the results would be expected to be fairly robust to differences in participants and protocol. If the findings are not robust to slight differences across studies, it seems unlikely they would generalise to clinical practice, in which there is natural variation in the clinical features of PwP, and the clinician approach to assessment.

Perhaps one of the reasons that results are not robust across studies, is that they relate to measures based on elements of bradykinesia, rather that the specific combination of measures in the bradykinesia definition. Abnormalities of speed alone or amplitude alone are not enough to reach the clinical definition of bradykinesia and yet it is these measures in isolation that are tested for group mean differences or tested for correlation with UPRDS-FT. This is partly because, in current definitions, the *presence* of bradykinesia requires a specific combination of deficits, but the *severity* of bradykinesia can be scored with only single deficits. It is difficult to understand the rationale for this discrepancy between chosen technology measures versus clinical definitions.

Bradykinesia is a complex, multifaceted phenomenon, based on a definition that includes nested 'and/or' statements. The current definition of bradykinesia is a formula, yet no studies recreate this formula with objective measures. It would be straightforward to report the combination of a speed measure plus an amplitude decrement measures and/or a measure of progressive hesitations/halts. It is unclear why this has never been pursued.

Machine learning studies tend to report high accuracy figures for classification into PwP/control or bands of UPDRS-FT scores, but they share one major limitation. This is the likelihood that the classification boundaries are based on a degree of 'over-fitting' to the specific data used in the study, and would not show the same performance figures if applied to PwP in general. Indeed, machine learning studies almost all test their classification algorithms against the same data that was used to train the algorithm, using methods such as 'leave one out' cross validation. In the absence of a very large number of participants, and testing of the algorithms on a completely new set of participants, the accuracy figures are likely to be artificially high due to overfitting.

There is also potentially a basic conceptual flaw in attempts to objectively quantify bradykinesia. That is that the idea of bradykinesia is a construction derived from expert consensus and clinician experience. It does not have its origins in empirical data. It has its origins in experienced clinicians attempting to describe what they have seen in the movement of PwP during clinical practice. As such, it is possible that bradykinesia as currently defined does not actually exist, at least in the exact form of its definition. That is not to deny that PwP have characteristic abnormalities of movement, visible to experienced clinicians, but just that those abnormalities may not exactly fit the current consensus definition of bradykinesia. The order of events is the opposite of many scientific definitions – for bradykinesia, the concept was invented many years before attempts at empirical measurement. In support of this, the definition itself has changed over time and across authors (as described in **Chapter 1**), suggesting it is not a fundamental natural phenomenon.

In conclusion, a large number of studies have used technology to measure movement during the finger tapping test in people with Parkinson's. The reported results vary considerably in significance, strength of correlation, and degree of accuracy. It is unclear whether these reports really demonstrate objective measurement or quantification of bradykinesia, partly because of methodological problems, and partly because the concept of bradykinesia itself is an invention of clinicians rather than something that exists exactly as defined.

Chapter 3, Parkinsonian hand or clinician's eye? Finger tap bradykinesia interrater reliability for 21 expert neurologists

Introduction

As described in **Chapter 1**, Parkinson's disease (PD) is a clinical diagnosis, and at the centre of this is the presence of bradykinesia: "*slowness of movement AND decrement in amplitude or speed (or progressive hesitations/halts) as movements are continued*" [7]. The gold standard test for bradykinesia presence and severity is a visual judgement made through the eye of an expert clinician, which almost always includes the finger tapping test, whereby an expert observes the patient repeatedly tapping their index finger against thumb "as quickly and as big as possible" [7, 8]. The finger tapping test is part of two standardised research rating scales: the Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale, MDS-UPDRS [8], **Table 1.1**, and the Modified Bradykinesia Rating Scale, MBRS [27], **Table 1.2**. In those scales, and in clinical practice, the observer judges three elements of finger tapping bradykinesia - speed, amplitude and rhythm. They are aggregated into a composite score in the MDS-UPDRS, but three separate scores in the MBRS.

Visual judgement as the gold standard evaluation for bradykinesia is problematic. Human assessment of movement is imprecise, with frequent disagreement amongst observers [95]. Bradykinesia is a complex, heterogeneous clinical sign that may be difficult to gauge accurately. That inaccuracy would mean that subtle changes of parkinsonism are difficult to measure, blunting the accuracy of clinical decisions, both for diagnosis and monitoring, and research outcomes.

It is noteworthy that a robust estimate of interrater reliability for finger tapping bradykinesia has not been published. There are several reasons for this. First, almost all studies have used very few (between 2 and 5) raters [27, 56, 96, 131, 155, 156], which is likely to be too few to assess the range of variability in clinician judgements. Second, most studies involved clinical raters applying the entire UPDRS motor examination to each participant [55, 96, 131, 155-

157], thus providing additional clinical information that influences the rater's judgement for any specific aspect of the examination. Henderson *et al.* [55] previously demonstrated this effect, showing that there was greater variation in rater scores when finger tapping was assessed in isolation (Kendall's W 0.5-0.6), rather than alongside other clinical assessments (Kendall's W >0.8). Third, most studies involved only people with PD, without any healthy control participants [27, 55, 56, 155-158]. This artificially avoids the difficult but important distinction between subtle bradykinesia and normal older age movement. Fourth, in some studies, Parkinson's medications are withheld prior to rating, thus exaggerating bradykinesia and making differences larger and therefore easier to detect [27, 55, 56, 156]. Fifth, only one [158] interrater reliability study has used the current MDS-UPDRS and only two have used MBRS [27, 56]. All other previous studies of interrater reliability for bradykinesia have used the older (now obsolete) version of UPDRS [55, 56, 96, 131, 155-157, 159], which has substantial differences in how the grades of bradykinesia severity are defined.

These methodological problems mean that we still do not know how well neurologists agree on such a central clinical sign, and the published figures for interrater reliability can vary widely for finger tapping bradykinesia. Cohen's κ of -0.07 (poor agreement or no agreement) [131, 160], κ of 0.47 (fair agreement) [155, 160], and Kendall's W of 0.87 (almost perfect agreement) [157, 160] have all been reported. We aim to address this by comparing 21 expert neurologists' bradykinesia ratings for finger tapping when no other information is given, in people with Parkinson's and also in people without a neurological diagnosis, with a statistical method appropriate for ordinal rating data.

Materials and Methods

The study was approved by the North of Scotland Research Ethics Committee, United Kingdom Health Research Authority (IRAS project ID 256116). Informed, written consent was obtained from all study participants.

Finger tapping video

39 people with idiopathic Parkinson's disease (PD) and 30 controls without a neurological diagnosis provided written consent. All PD participants had previously been diagnosed by a

movement disorder specialist neurologist at Leeds Teaching Hospitals NHS Trust, United Kingdom, according to Movement Disorder Society clinical diagnostic criteria [7]. PD participants were subjectively and objectively in the 'on' state at the time of participation (no medications were withheld). One investigator, SW, graded Hoehn and Yahr stage for each participant, and also later scored the presence / absence of visible tremor in each video (but did not score any video for bradykinesia). Healthy controls were recruited from the companions of patients and hospital/university staff. They had no history of Parkinson's or other neurological diagnosis, and were not taking any medication that could induce parkinsonism.

Participants rested their elbow on a chair arm with the forearm lifted at 45°. In accordance with MDS-UPDRS instructions, each participant was instructed to tap their index finger and thumb together "as quickly and as big as possible" with each hand examined separately. The participants tapped for just over 10 seconds, because the MDS-UPDRS specifies 10 taps while the MBRS specifies 10 seconds [8, 27].

We recorded videos of each hand during the task using a standard smartphone (iPhone SE) placed on a tripod (60 frames per second, 1920x1080 px) under ambient lighting. Only the hand and part of the forearm were within the video frame. The distance from camera to hand was approximately 1m and digits 1 and 2 were closest to the camera.

One video was discarded because the hand moved outside the video frame, making 137 videos: 77 Parkinson's hands and 60 control hands. Each video was edited to contain 1 second prior to tapping onset and 10 seconds of finger tapping.

Clinical rating

We invited 21 consultant neurologists that specialise in movement disorders, from a range of clinics in the United Kingdom, to each rate 30 videos of finger tapping. Python [161] was used to select 30 random videos for each neurologist from the total set of 137 videos – for each clinician the list of all videos was shuffled randomly and the first 30 used. Each video was rated according to the MDS-UPDRS Item 3.4 Finger Tapping [8] (first ten taps) and the MBRS [27] (the full 10 seconds of tapping), **Table 1.1** and **Table 1.2**. The neurologists undertook the task independently, at separate locations, on their own computer screen, and were blinded to both PD / control status and to each other's scores.

Inspired by informal comments made by the first two raters, we added an additional question for the subsequent 19 neurologists - asking them to judge whether the hand was most likely to be from a control or PD participant. This was in recognition that an experienced clinician may form an overall, subjective impression about whether the tapping appears Parkinsonian or not, that is not necessarily strictly based on bradykinesia criteria.

Outcomes

The primary outcome is the interrater reliability for MDS-UPDRS finger tapping scores, reported as the intraclass correlation coefficient (ICC), which was the basis of the statistical power calculations. The secondary outcomes are: correlation coefficients describing the relationship between MDS-UPDRS score and each of the three MBRS score components, the proportions of healthy controls rated as bradykinesia by MBRS sub-score combination, and the accuracy of clinicians in judging PD from controls.

Statistical analysis

Interrater reliability reflects the variation between more than one rater measuring the same group of participants [162]. We report ICCs for both *absolute agreement* and *consistency*. Absolute agreement concerns the degree to which one rater's score (x) is exactly equal to another's (y), whereas consistency concerns the degree to which x can be related to y plus a systematic error (x + c).

For each ICC, we calculate scores using a standard linear model, which assumes the underlying normal distribution of MDS-UPDRS scores, and also a more sophisticated novel approached based upon cumulative linked mixed models (CLMMs), which is more appropriate for dealing with ordinal data. The normal distribution assumption of the first model is clearly incorrect, but allows direct comparison to previous research. Both approaches are two-way random effects models, where each item is assessed by the same set of raters randomly selected from a larger population of raters. Note that, typically, a two-way random effects model would have all raters viewing all videos, which is impractical for our scenario. Our approach is equivalent to taking a random sample of this “ideal” complete dataset, which gives unbiased estimates but enlarged confidence intervals.

The random effects models consist of a random effect for video number (capturing the tendency of a video to be scored higher or lower than expected), a random effect for rater number (to capture the tendency of a rater to under-/over-rate videos), a fixed effect for whether the video is of a patient or control participant to give a baseline score in each case, and an intercept term. If σ_v^2 denotes the variance of the random effect for video number, σ_r^2 is the variance of the random effect for rater, and σ_ϵ^2 is the variance of the residual error then the agreement ICC is calculated as follows.

$$\frac{\sigma_v^2}{\sigma_v^2 + \sigma_r^2 + \sigma_\epsilon^2}$$

Meanwhile the consistency ICC is calculated as follows.

$$\frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2}$$

We fit two models to the data for calculating the ICC. The first uses a normal approximation to the ordinal score as in previous work. Our second model keeps the dependent variable ordinal using a cumulative linked mixed model (CLMM) – essentially fitting a latent normal model with the addition of “cut-points” which split the latent normal distribution into segments corresponding to the dependent ordinal variable [163].

Whilst this latter CLMM readily gives the variance of the random effects for video numbers and raters, it is not initially clear how to define the residuals, which are required to calculate the ICC. In effect we need to define the “optimal” value in the latent space for each level that the ordinal variable can take. We took the following approach: after fitting the latent normal distribution and cut-points the optimal points were defined as the median of each segment of the normal distribution (calculated using Monte Carlo). With these points defined, the residual can be calculated using the latent value of the fitted model on each data point and the corresponding optimal values.

The study power calculation was done via simulation using the normal approximation to the ordinal variable, based on pilot data with two raters. Based on recruiting 20 raters and covering a variety of different strength ICC values, we determined that giving 30 random videos to each rater allows us to calculate the ICC to within 0.05 in 95% of trials and to within 0.03 in 80% of trials. Models were fitted using the R libraries 'glmer' and 'clmm', whilst power calculations were done using the Python library 'statsmodels' [161].

Secondary analysis consisted of calculating the three Spearman correlation coefficients of the relationship between the median MDS-UPDRS score across all raters, with the each of the median MBRS speed score, amplitude score, and rhythm scores.

Results

Expert neurologists' rating of finger tap bradykinesia in people with Parkinson's and controls

The age, gender and Hoehn and Yahr scores for the participants are given in **Table 3.1**. The median number of raters per video was 5 (range 1 to 12, interquartile range 3 to 7). In the random selection of 30 videos per rater, 4 videos from the total of 137 were not allocated to any rater, so that the total number of unique hand videos rated was 133. A total of 630 video ratings were made (21 raters, 30 videos each): 325 of these were ratings of PD videos, and 305 ratings of healthy control videos.

The distribution of MDS-UPDRS finger tapping scores for PD and control videos are shown in **Figure 3.1**. 53% of control participant videos were given an MDS-UPDRS finger tapping score greater than 0. The distribution of MBRS scores for finger tapping speed, amplitude and rhythm are shown in **Figure 3.2**. Across both rating scales, scores of grade 1 ('slight' impairment by MDS-UPDRS, 'mild' impairment by MBRS) were similarly common in both control videos and PD videos. The proportion of videos scored grade 1 by MDS-UPDRS was 26% in PD and 34% in healthy controls, while the proportions of videos scored grade 1 for MBRS speed, amplitude and rhythm were 40%, 22%, 31% respectively in PD, compared with 31%, 21%, 27% respectively in controls.

	People with Parkinson's	**Healthy control participants**
Age (Std. Dev.) yrs	68 (9.6)	59 (19.4)
Male/Female	47/26	22/38
Median years since diagnosis	4	n/a
Median H&Y [IQR]	2 [1,3]	n/a
H&Y = 1	32	
H&Y = 1.5	2	
H&Y = 2	12	
H&Y = 2.5	4	
H&Y = 3	19	
H&Y = 4	4	
H&Y = 5	0	
	People with Parkinson's	**Healthy control participants**
Impaired speed	77%	43%
Impaired rhythm	72%	35%
Impaired amplitude	70%	30%
Impaired speed and rhythm	62%	19%
Impaired speed and amplitude	61%	19%
Bradykinesia (Impaired speed + impaired rhythm and/or impaired amplitude)	**64%**	**24%**

Table 3.1. Participant (hand video) characteristics, and proportion of finger tapping videos rated as bradykinesia and its components by MBRS subscores. Hand video characteristics are split by Parkinson's hands (*n* = 73) and control hands (*n* = 60). H&Y: modified Hoehn and Yahr scale [164]. IQR: Interquartile Range.

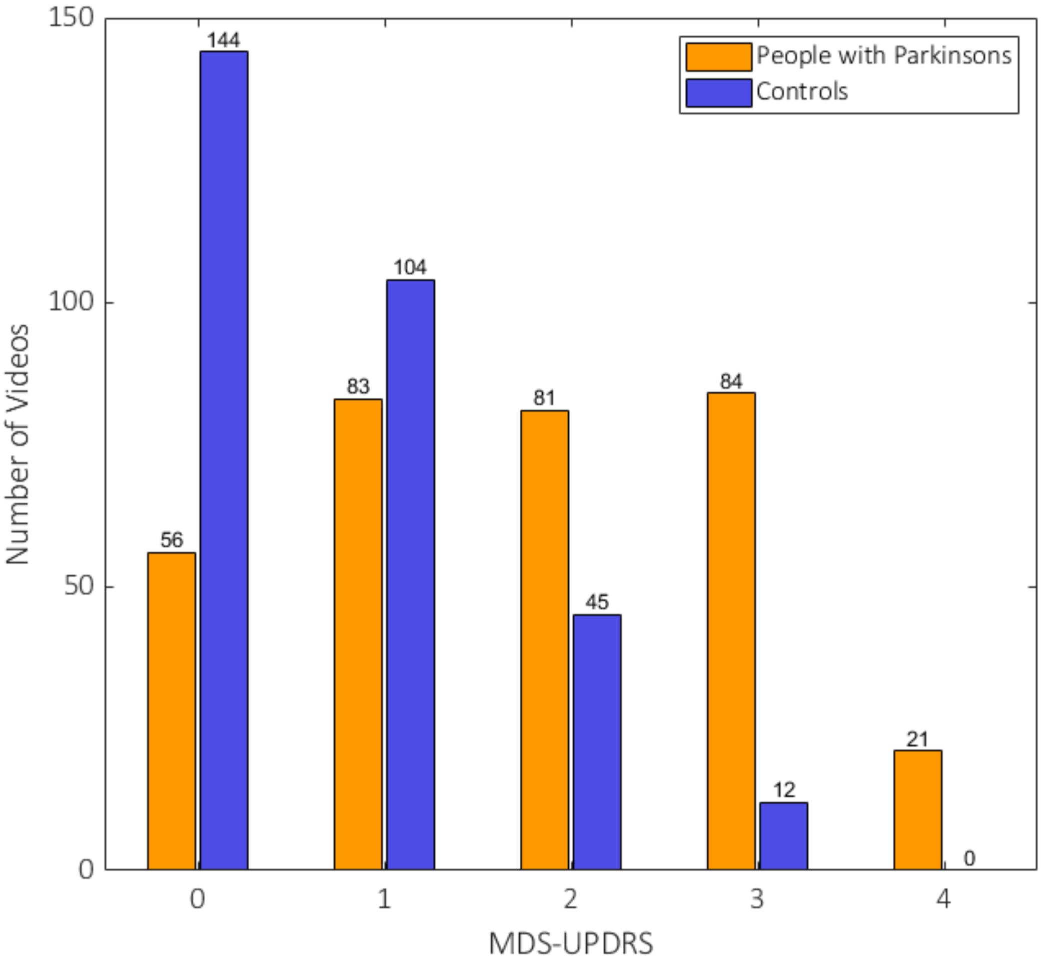


Figure 3.1. Histogram showing the distribution of MDS-UPDRS finger tapping scores for all ratings of the hands of people with Parkinson's (orange bars) and control participants (blue bars).

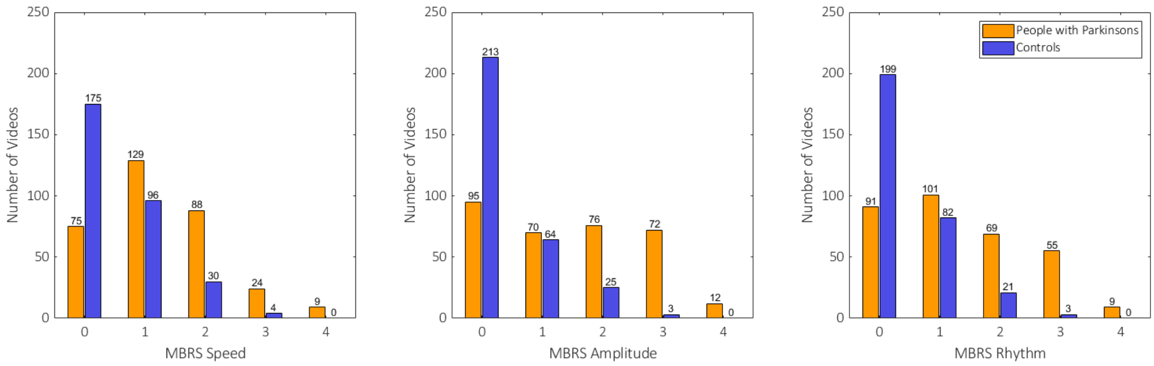


Figure 3.2. Histograms showing the distribution of MBRS finger tapping scores for all ratings of the hands of people with Parkinson's (orange bars) and control participants (blue bars).

Bradykinesia is defined as slowness of movement AND decrement in amplitude or speed (or progressive hesitations/halts) as movements are continued [7]. Therefore, the MBRS subscores for finger tapping can be used to classify tapping as bradykinesia if a rater scores a video >0 for speed and also >0 for amplitude and/or rhythm. **Table 3.1** shows the proportions of videos in PD and controls (respectively) with impaired speed, rhythm, and amplitude, as well as combinations of those deficits, and the specific combination that meets the definition of bradykinesia. Among PD videos, 77% were rated as slow, and 64% were rated as bradykinesia by MBRS (>0 for speed and >0 for one or more of amplitude or rhythm). Among videos of control participants, 43% were rated as slow, and 24% were rated as bradykinesia by MBRS (>0 for speed and >0 for one or more of amplitude or rhythm). Thus, one in four control participant hand videos were rated as bradykinesia by MBRS.

Interrater reliability for finger tapping bradykinesia

The intraclass correlation coefficient (ICC) for MDS-UPDRS rating of finger tapping bradykinesia for exact agreement was 0.53 using the normal model ('fair' [165] or 'moderate' [162]) and 0.65 using the cumulative linked mixed model ('good' [165] or 'moderate' [162]). The ICC for consistency (ratings related to each other with a systematic error) was 0.58 using the normal model ('fair' [165] or 'moderate' [162]), and 0.78 using the cumulative linked mixed model ('good' [165] or 'moderate' [162]).

To assess model discrimination for the CLMM, we investigated the predicted values with the original ratings. The CLMM predicts the correct MDS-UPDRS score with 70% accuracy and is accurate to within one point on the five-point MDS-UPDRS finger tapping scale 98% of the time.

Figure 3.3 shows the variation in clinical ratings. Each point is an individual clinical rating of a video: the x-axis orders the videos by CLMM random effect size, and the y-axis is the clinical MDS-UPDRS rating. The values are jittered in the y-axis for visual clarity. It demonstrates the considerable variation in movement disorder specialist judgement of individual videos, with disagreement common.

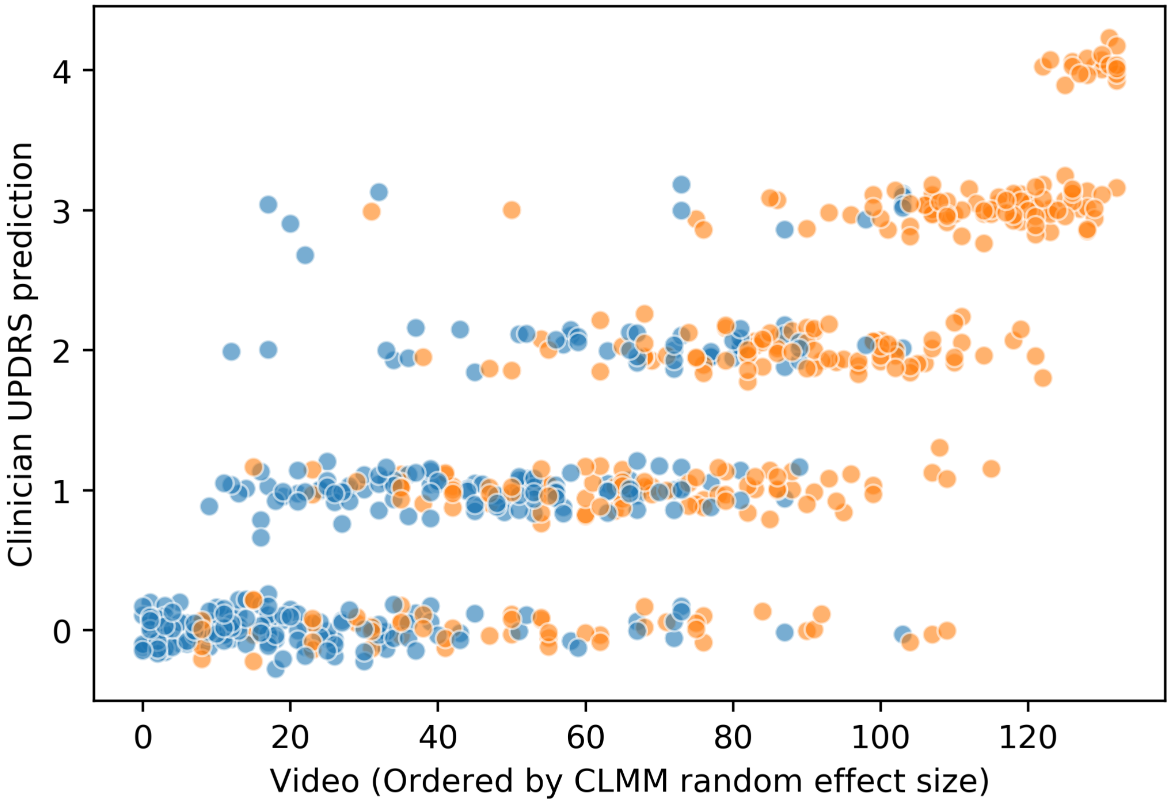


Figure 3.3. Distribution of movement disorder specialist MDS-UPDRS ratings for finger tapping videos. Each circle represents a video rating (jitter applied to aid visualisation). Orange circles: people with Parkinson's. Blue circles: healthy control participants. Videos are ordered on the x-axis by the video random effect size according to cumulative linked mixed model, CLMM (i.e. by severity of bradykinesia). It can be seen that there is considerable variation in clinician ratings.

Correlations between finger tapping MDS-UPDRS and individual MBRS elements

The Spearman correlation coefficients for MDS-UPDRS finger tapping scores and each of the MBRS subcomponent scores were R=0.77 for speed, R=0.78 for amplitude, and R=0.68 for rhythm, **Figure 3.4**.

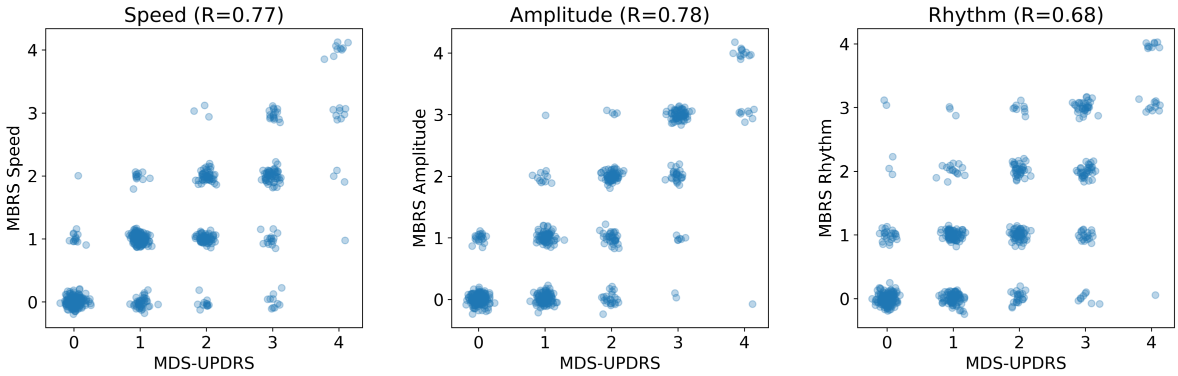


Figure 3.4. Correlation between MDS-UPDRS scores and MBRS subscores (speed, amplitude, rhythm). Each blue circle is one rater and one video. Jitter added to aid visualization. Correlation strengths are similar, suggesting raters do not prioritise any particular bradykinesia subcomponent in assigning an overall MDS-UPDRS score.

Neurologists' judgement of whether finger tapping video shows a person with Parkinson's or control

The movement disorder specialists correctly judged PD or control status in 70% (400 of 570) videos. The median number of correct judgements was 20/30 (67%), with a range from 17/30 to 27/30, interquartile range 18.75 to 23.5 (out of 30).

Of those videos judged to show a PD hand, 77% were formally rated as showing bradykinesia by the relevant MBRS subscore combination. Of the *correctly identified* PD videos, 84% were scored as bradykinesia. In other words, the movement disorders specialists' overall perception of PD or control was not strictly related to the presence or absence of bradykinesia by MBRS subscore combination. Nor was it simply explained by visible tremor. Of 36 PD video ratings not rated as bradykinesia but judged to show a PD hand, only 8 had visible tremor in the video (while 9/69 PD videos rated no bradykinesia and judged to show a control had visible tremor). Among videos correctly judged to show a control hand, 5% were formally judged as showing bradykinesia. Of the *correct* control judgements 3% were scored as bradykinesia.

Discussion

Our results demonstrate that even expert neurologists frequently disagree about the level of bradykinesia on finger tapping, despite clinical examination representing the gold standard for determining the presence and degree of bradykinesia [7, 8]. The 21 movement disorder specialists showed only 'moderate' agreement [162] for MDS-UPDRS finger tapping ratings (ICC=0.53, CLMM-ICC=0.65). Furthermore, the same movement disorder specialists classified one in four healthy control participants as showing bradykinesia on finger tapping (using MBRS sub-scores to match the definition of bradykinesia), and the proportions of participants showing slight or mild abnormalities on MDS-UPDRS and MBRS was similar in PD and control videos. This suggests that finger-tapping bradykinesia is a non-specific sign and overlaps with changes in movement associated with normal ageing, at least when mild. It is perhaps unsurprising that bradykinesia is difficult to judge. It is a heterogeneous clinical sign, and human vision cannot accurately measure and compare movement speed, amplitude and rhythm in isolation, much less in simultaneous combination.

Our findings are particularly robust because they are based on a larger number of raters (21) and unique videos (137) than previous studies. Each neurologist rated 30 videos and the median number of raters per video was 5, but these numbers were based on statistical power calculations, and the random distribution of videos to raters mean that variation among the whole group is well characterised. Another strength of this study is the use of a cumulative linked mixed model, respecting the ordinal nature of MDS-UPDRS scores, a consideration that has been neglected in previous research. Furthermore, we not only reported MDS-UPDRS finger tap ratings, but also MBRS ratings, which separately score each of tap speed, amplitude and rhythm. In contrast to a 2011 study, which found that clinicians weighted amplitude and rhythm more than speed in UPDRS bradykinesia scores [27], we found strong correlations for all MBRS subscores with MDS-UPDRS (0.68-0.78), with rhythm the weakest of the three, suggesting that clinicians do not favour any particular subcomponent of bradykinesia in finger tapping judgements. We also reported consistency ICCs, which were a little higher than agreement results (ICC=0.58, CLMM-ICC=0.73), but in a five-point scale, consistent inter-rater variation (a consistent difference between raters) is of little clinical relevance compared with absolute rater agreement. For example, two raters who disagreed by 2 or 3 points on every video would nevertheless show very high consistency ICC if that 2 or 3 point difference in ratings was consistently present (and in the same direction) across all the videos.

A previous study of a UPDRS 'teaching tape' supports the idea that finger tapping bradykinesia is difficult to judge [95]. 226 raters were tested in their UPDRS motor scores for 4 people with PD (using video recordings). A 'pass' in this test was defined as a score within the 95% confidence interval of 3 international Parkinson's disease experts for each case. Only 54.6% of raters 'passed' the 4 cases, and of those that 'failed' first time, 70.6% failed finger tapping rating.

Previous studies of finger tapping interrater reliability by UPDRS grading have reported Kendall's W 0.84 and 0.87 [157], weighted κ of 0.53 to 0.71 [96], 0.72 to 0.86 [156], κ of 0.47, 0.44, -0.07 [131, 155], and Kendall's τ of 0.88 and 0.84 [158], while MBRS raters showed Pearson correlations of 0.51, 0.77 and 0.69 respectively [27]. It is difficult to draw conclusions from those results because of methodological limitations that include low numbers of raters and/or people with PD [27, 56, 96, 131, 155, 156, 158] (including non-overlapping subsets of raters [157]), the absence of 'healthy control' participants [27, 55, 56, 155-158], participants 'off' their usual medication [27, 55, 56, 156, 158], statistical methods inappropriate for ordinal data [27], measures of simple correlation rather than agreement [158], and raters gaining additional information from the entire UPDRS or UPDRS motor exam [55, 96, 131, 155-157].

It could perhaps be argued that the influence of a broader UPDRS assessment upon finger tapping scores is appropriate, reflecting clinical practice, in which finger tapping would never be tested in isolation. However, busy routine clinics do not involve enough time for the complete UPDRS (a "vast instrument" [55]). Furthermore, limb bradykinesia must be documented to establish a PD diagnosis, although bradykinesia also occurs in the face, voice, and axial/gait domains [7]. In addition, UPDRS bradykinesia items are commonly analysed as a standalone 'bradykinesia' endpoint in trials [27], or used as a gold standard for demonstrating that technological devices 'quantify' bradykinesia [12-47]. Most fundamentally, finger tapping bradykinesia is presented in the literature as a measure of a specific phenomenon with a specific definition. It is not defined as a surrogate for an overall impression. If the latter is to some extent true, then it becomes less clear exactly what bradykinesia actually is [69], and less clear that movement disorder specialists are able to define and measure this “cardinal manifestation” [7] of PD.

In our results, one in four control videos were rated as showing finger tapping bradykinesia (using MBRS subscores). This is consistent with a previous study in which three trained nurses and one movement disorder specialist rated older people with no clinical Parkinson’s according to a modified UPDRS motor score [96]. They gave 74 out of 75 participants a score greater than 0 (mean score 13.4 out of 127). Of course, the MDS diagnostic criteria for PD are not based on bradykinesia alone, and instead require a combination of clinical features to be present or absent to diagnose PD [7]. However, to some extent this only amplifies the challenge for clinician reliability, because other clinical features such as tremor are also non-specific, and there is considerable evidence that the overall diagnostic assessment of PD is difficult, with less-than-ideal sensitivity and specificity. This includes misdiagnosis rates of PD versus Essential tremor of one in three [91], as well high false positive (17.4-26.1%) and false negative (6.7-20%) rates for the diagnosis of PD based on video examinations of people with tremor [92]. One postmortem study showed misdiagnosis of PD in 24 out of 100 cases [132].

We asked the clinicians to judge whether the hand in the video was most likely to be that of a person with PD or a control. Of those videos guessed to show the tapping of a person with PD, only 77% were judged to show bradykinesia by the appropriate combination of MBRS subscores. This suggests the possibility that clinicians are forming an overall impression of finger tapping that does not purely follow the formal definition of bradykinesia: a gestalt perception or intuitive pattern recognition of finger tapping normality / abnormality beyond the presence or absence of bradykinesia as defined by formal criteria [140-142]. In support of this idea, a clinicopathological study found that experienced movement disorder specialists showed a higher accuracy than claimed for most clinical diagnostic criteria, for the diagnostic distinction of different forms of parkinsonism. The authors state that these experts, “may be using a method of pattern recognition for diagnosis that goes beyond any formal set of diagnostic criteria” [166].

In conclusion, a classic sign of a cardinal clinical feature of a common neurological disease - finger tapping bradykinesia - is not easy to reliably see, even for expert eyes. Our findings suggest that bradykinesia is to some extent a phenomenon present in the eye of the clinician rather than simply in the hand of the person with Parkinson’s disease.

Chapter 4, Computer vision to measure finger tapping bradykinesia in video

Introduction

The phenomenon of bradykinesia is considered fundamental to Parkinson's diagnosis and assessment, but relies on the subjective visual judgement of a small group of experts. As **Chapter 3** showed, the human judgement of bradykinesia is difficult, with only moderate agreement between experts, and also a significant proportion of control videos judged to show bradykinesia. Previous attempts to use technology to automate bradykinesia assessment have frequently relied on specific hardware or patient interaction with a specific app, as described in **Chapter 2**.

The primary aim of the work in this chapter is to provide proof-of-concept that the assessment of bradykinesia can be automated using simple camera input, negating the impact of inter-rater variability and providing easily accessible clinical decision support. I describe two computer-vision based approaches to this, using video of hands performing the finger tapping test for bradykinesia.

The first approach – Study 1 – derives a measure of whole-hand movement over time from the video, and derives several pertinent features from this time series, chosen to reflect the subcomponents of bradykinesia. These features are used with machine learning techniques to classify tapping videos into two levels of bradykinesia: (a) no or slight bradykinesia, versus (b) mild to severe bradykinesia, and also to classify videos as Parkinson's versus control.

The second approach – Study 2 – uses a deep learning method to track the positions of key landmarks on the hand, allowing the relative distance between finger and thumb to be measured over time. From that time series, several features reflecting bradykinesia subcomponents are again derived. Each of these features is correlated against a large group of clinical ratings for the tapping videos

Methods

Ethical review

The work was approved by the North of Scotland Research Ethics Committee, United Kingdom Health Research Authority (IRAS project ID 256116). All participants gave written, informed consent.

Participants and video recording

Two groups of participants were recruited: people with idiopathic Parkinson's disease and control participants who did not have any neurological diagnosis (or medication) with the potential to affect movement.

Participants with idiopathic Parkinson's disease had been previously diagnosed by a movement disorder specialist neurologist at Leeds Teaching Hospitals NHS Trust, United Kingdom, according to Movement Disorder Society clinical diagnostic criteria [7]. They were subjectively and objectively in the 'on' state at the time of video recording (no medication was withheld prior to recording). Recruitment did not exclude patients with postural hand tremor or dystonia. The control participants were recruited from the companions of patients, or hospital/university staff.

Video recordings were made of participants' hands performing the finger tapping test for bradykinesia. Each hand (left, right) was filmed individually. In study 1, 70 finger tapping videos were used, from 20 participants with Parkinson's (40 hands) and 15 control participants (30 hands). In study 2, 137 videos were used, as the original participant group was extended to 39 Parkinson's participants (77 hands) and 30 controls (60 hands). The Parkinson's group in study 2 was 77 hands not 78, because in one video the hand moved out of the video frame, and the video was discarded.

Participants rested their elbow on a chair arm with the forearm lifted to a 45 degree angle. The hand was free to move as per the protocol of the Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Item 3.4 Finger Tapping [8]. Only the hand/forearm was within the video frame. The distance from camera to hand was approximately 1m, but not tightly defined. Digits 1 and 2 were closest to the camera. No specific instructions were given for the position of digits 3 to 5, and participants were free to position these digits as they preferred, although the researcher gave a brief demonstration in which digits 3 to 5 were extended.

A smartphone, placed on a tripod, was used to record standard video (60 frames per second, 1920x1080 pixel resolution), with only ambient lighting. Participants were instructed to tap their index finger and thumb together “as quickly and as big as possible” for 15s [8].

Methods: Study 1

The degree of bradykinesia in each video was independently rated by two consultant neurologists with a special interest in Parkinson’s, according to the section 3.4 of the MDS-UPDRS scale (UPDRS-FT), **Table 1.1** [8]. The raters were blinded to patient/control group.

For both groups, the correlation between UPDRS-FT scores from the right and left hand for an individual participant was very low (Patients k = 0.17, 95% CI: − 0.18 to 0.47, Controls k = 0.18, 95%, CI: − 0.07 to 0.41). Consequently, we treated videos from each hand as independent samples.

Data processing

A schematic of the data processing framework is presented in **Figure 4.1**.

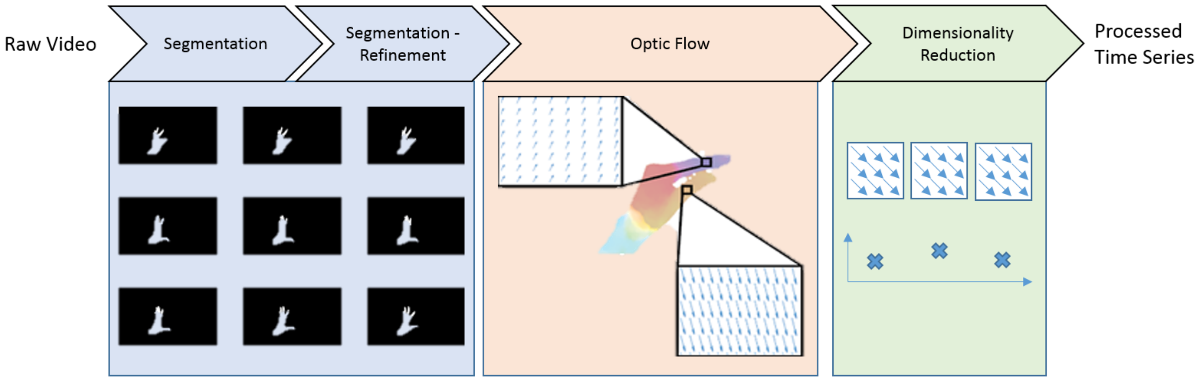


Figure 4.1. Illustration of the data processing in which raw video is converted to an anonymous 1D time series. Raw video is first segmented using a convolutional neural network. The segmentation is refined using the grabcut method. Frame-by-frame movement of the hand is extracted using optical flow. The optical flow field is then reduced so that the magnitude of movement between two frames is summarised by a single value.

Initially, the video frames were segmented to the pixels corresponding to a participant's hand. Traditional skin colour methods were unsuitable, given the uncontrolled lighting conditions used. Instead, the hand regions of interest were first detected using a convolutional neural network, originally proposed by Bambach *et al.* [167]. The detector is based on a MobileNet-V2 mode architecture and the single shot multi-box approach using the TensorFlow Object Detection API [168, 169]. This architecture uses depth-wise separable convolutions that reduce the computing overhead, so that it can be used in mobile devices. We trained our model using manual annotation of the hand region in 500 randomly selected frames from our dataset.

The output of the model was refined using a secondary pixel-level segmentation to remove erroneous background pixels. We used the 'GrabCut' method [170], which iteratively updates two Gaussian Mixture Models representing the background and foreground. We set two mixture components to model the foreground colours and 3 mixture components for the background colours.

The segmented frames were then converted into an optical flow field [108]. In such a field, each position corresponds to the vector pixel movement of a point object between two sequential frames. The magnitude of the vector thus represents the instantaneous speed of a point (in pixels/frame). We sum the magnitude at each point in the region of interest to obtain a metric of overall hand movement.

Optical flow magnitude is affected by camera distance and hand size (as well as actual movement), so to convert optical flow magnitude into true hand velocity, we scale the magnitude by the number of pixels in the hand region of interest, such that our metric *Mt* is:

$$M_t = \frac{\sum_j^H \sum_i^W b_{\mathrm{ij}} \sqrt{u_{\mathrm{ij}}^2 + v_{\mathrm{ij}}^2}}{\sum_j^H \sum_i^W b_{\mathrm{ij}}},$$

where *H* and *W* are the height and width of the optical flow field, *u* and *v* are the horizontal and vertical components of the flow, and *b* is the pixel mask obtained from the image segmentation. By evaluating *Mt* over a sequence of video frames we produce a 1D signal over time. Examples of the signal are shown in **Figure 4.2**.

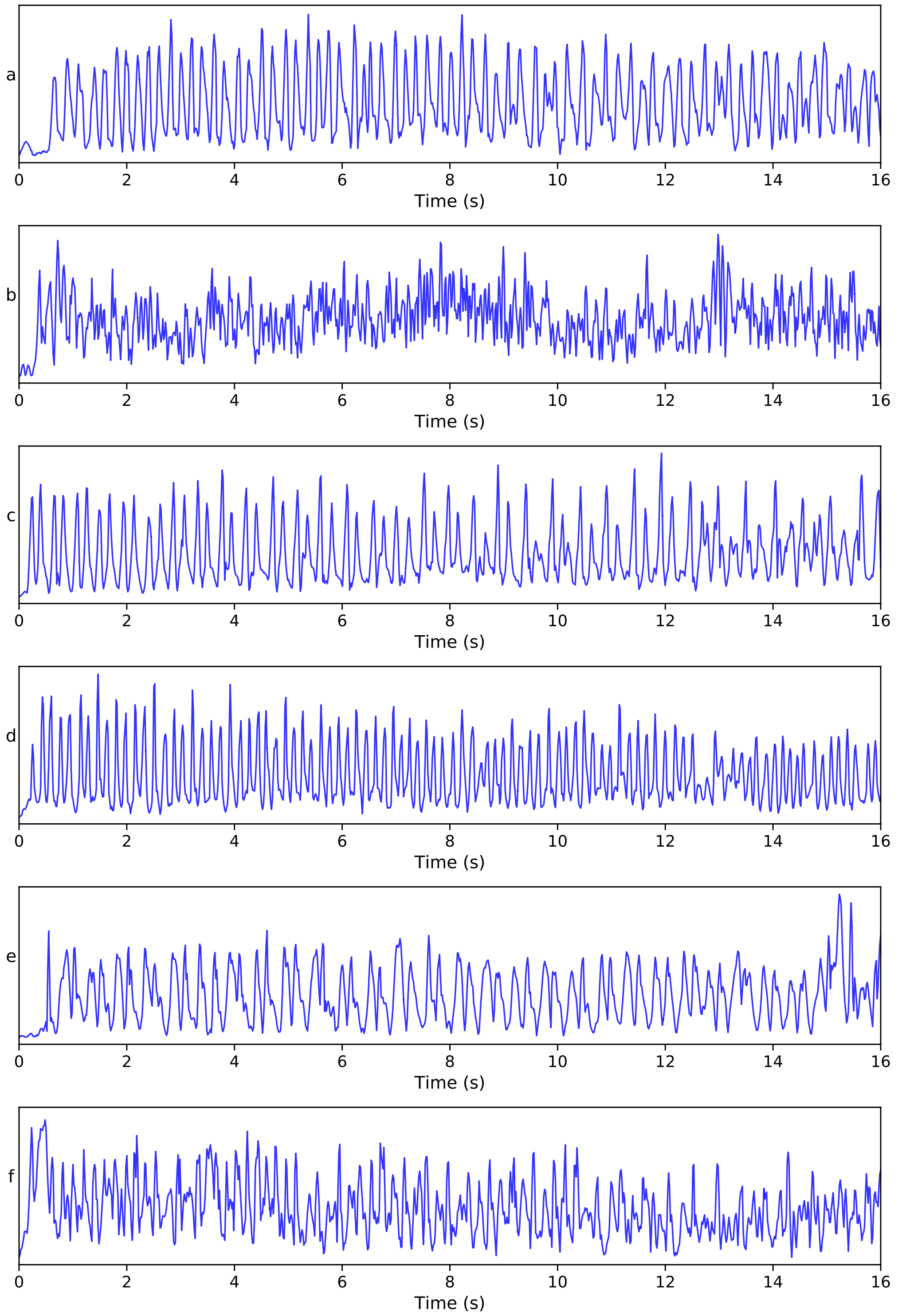


Figure 4.2. Examples of the optical flow magnitude time series, plots (c)-(f) are discussed in the section headed ‘misclassified UPDRS-FT categories’. **(a)** – no bradykinesia (UPDRS-FT = 0). **(b)** – severe bradykinesia (UPDRS-FT = 4). **(c)** – UPDRS-FT = 0–1 misclassified as UPDRS- FT=2–4, close to decision boundary. **(d)** – UPDRS-FT = 2–4 misclassified as UPDRS- FT=0–1, close to decision boundary. **(e)** – UPDRS-FT = 0–1 misclassified as UPDRS- FT=2–4, far from decision boundary. **(f)** – UPDRS-FT = 2–4 misclassified as UPDRS- FT = 0–1, far from decision boundary.

Feature extraction

Candidate features were derived from the 1D signal via clinical knowledge and visual inspection. In particular, we derived a set of features that described the frequency, amplitude, and tap-to-tap variability, to reflect the MDS-UPDRS finger tapping assessment criteria, as follows.

Frequency: **Tapping frequency** was estimated as the frequency corresponding to the maximal amplitude peak in the fast Fourier transform (FFT) spectrum. This assumes that the finger tapping motion corresponds to the greatest movement (and thus energy) between frames and that other movements, such as tremor, have smaller magnitude.

Amplitude: **Energy spectral density** was calculated as the squared integral of the FFT spectrum, a measure that would be expected to increase with the amplitude of tapping. In addition, we assumed that bradykinesia movement is distinctive in some frequency bands. Therefore, the energy spectral density is separated into six non-overlapping equal frequency bands ranging from 0 to 18.36 Hz with bandwidth interval 3.06 Hz. The upper frequency threshold was selected heuristically to avoid having multiple uninformative zero-energy frequency bins. The threshold represents the frequency up to which, on average, 99% of the signal energy is contained.

Variability: Two variability features were derived using the peaks of the optical flow waveform. Peaks were calculated via the MATLAB function *findpeaks* with zero minimum peak prominence [171]. Peaks were then classified as *maxima* or *minima* by fitting a 1D Gaussian mixture model with two clusters to the peak amplitude values. We then defined:

Jitter: We hypothesise that there are differences between the hand closing and hand opening motions. From visual inspection, we observed differences in higher frequency movement between the signal *maxima* and *minima* – troughs in the signal appeared more jittery than the peaks. To quantify the jitter we include the ratio of number of *maxima* to number of *minima* over the entire time series as a predictor.

Peak-to-peak variability: was calculated as the standard deviation of the time between *maxima* peaks. This feature models variation in tapping frequency across the time series and may be considered analogous to the standard deviation of RR intervals (SDRR) for ECG signals [172].

Classification

We performed binary classification using Naïve Bayes (NB), logistic regression (LR), and both linear and RBF-based support vector machines (SVM-L and SVM-R, respectively) [173] to predict two outcomes: (1) a UPDRS-FT score > 1, and (2) clinical diagnosis of Parkinson's disease (previous clinical diagnosis by a consultant neurologist). Where there was disagreement in rater UPDRS-FT scores, the higher score was selected for training of the models.

Given the relatively small number of samples in the dataset we began by reducing the feature space into two dimensions using principal component analysis. Indeed, preliminary work fitting models with all 10 features led to significant overfitting. We then explore the effect of analysing up to 5 principal components, to look for any additional gain in accuracy.

The NB model was chosen as a simple baseline classifier providing a sensible lower bound for performance.

LR provides a linear separation of the data points and this simplicity may lead to lower generalisation error. We incorporated ridge (L2) regularisation with strength determined via a grid search of 100 log- spaced values in the interval [1e − 4, 1e+4] to minimise 10-fold cross-validation accuracy loss.

The SVM-L model optimises a different cost function than the LR model and therefore gives a different linear separation of the classes. Meanwhile, the SVM-R model has the ability to model nonlinear decision boundaries. The slack and (for SVM-R) kernel scaling hyper-parameters were again estimated using a grid search to minimise 10-fold cross-validation accuracy loss. The grid search consisted of 100 log-spaced values in the intervals [1e+0, 1e+3] and [1e+0, 1e+5], respectively.

We report the training accuracy and AUC score for each model with two principal components, and for 3–5 components. We used permutation tests (α = 0.05) on the variant obtaining highest accuracy to assess whether classifiers had meaningful predictive ability [174].

Due to the relatively small size of our pilot data we estimate the out-of-sample test accuracy, sensitivity, and specificity of each model by reporting the mean value of leave-one-out cross-validation (LOO-CV). Hyperparameters were preset according to outputs of the 10-fold cross-validation procedure described above.

We also investigate the contribution of each feature to the principal component analysis, to investigate the most discriminative features of the timeseries and compare with other research on this topic.

Finally, a visual inspection of the raw videos underlying the timeseries that were misclassified by the model with highest LOOCV accuracy was performed by two neurology clinicians. Analyses were performed using MATLAB 2017b and the scikit-learn and TensorFlow packages for Python 3 [171, 175].

Methods: Study 2

Clinical rating

22 consultant neurologists specialising in movement disorders (United Kingdom) were asked to rate 30 tapping videos each, selected at random from the set of 137 videos. The median number of raters per video was 5 (range 1 to 12, interquartile range 3 to 7). The raters were blinded to patient / healthy control status and each other's scores. Each video was rated according to both the MDS-UPDRS Item 3.4 Finger Tapping [8], and the Modified Bradykinesia Rating Scale (MBRS) [27, 56]. The MDS-UPDRS requires the rater to amalgamate judgments of finger tapping speed, amplitude, and rhythm into a single composite score (**Table 1.1**). In contrast, the MBRS is comprised of three separate scores for speed, amplitude, and rhythm (**Table 1.2**). The modal clinician rating for each video was used for correlation with computer measures (see below). The inter-rater reliability for this group of raters and videos is described in **Chapter 3**.

DeepLabCut video tracking

DeepLabCut tracks the geometrical configuration of multiple body parts in video, without a requirement to wear markers [116, 117]. It uses transfer learning, with feature detectors based on deep neural networks that have been pretrained on ImageNet, a massive object recognition dataset. This means that DeepLabCut is able to accurately recognise and track body parts with minimal training data [117].

The set of 137 finger tapping videos was processed by DeepLabCut. We localised (labelled) six distinct points on the tapping hand in 20 frames selected by k-means clustering from each 660 frame video: thumb tip; index finger tip; thumb metacarpophalangeal (MCP) joint; index finger MCP joint; middle finger tip; dorsal wrist/proximal dorsal hand. The deep neural network architecture of DeepLabCut was then trained using these points to predict their localisation in the remaining 97% of (unlabelled) video frames (1030000 training iterations, ResNet 50). This created video pixel coordinates for finger tip and thumb tip throughout each video.

The accuracy of DeepLabCut to track the hand localisation points (including finger tip and thumb tip) was assessed using the 'evaluate network' function within DeepLabCut. This function computes the Euclidean error between the manual labels and the ones predicted by DeepLabCut averaged over the hand locations and test images (mean absolute error, proportional to the average root mean square error) [116].

Signal processing

Video pixel coordinates for the labels produced by DeepLabCut were used to calculate the pixel distance between index finger tip and thumb tip. A Savitzky-Golay filter was applied to the resulting time series, which removed large, sudden transient label 'jumps' caused by DeepLabCut mislabelling (i.e. large-distance, physiologically impossible label movement across one pair of video frames). The pixel number distances were standardised across all videos, by using the maximum opening distance between finger and thumb tip and normalising this to a value of 1 (all values were divided by the maximal value in the corresponding time series).

Three features of the resulting finger tip to thumb tip distance time series were calculated to reflect the clinical features of finger tapping speed, amplitude and rhythm, **Figure 4.3**. A measure of speed was calculated as the mean rate of change of the normalised distance between finger and thumb tip over time. A measure of amplitude variability was calculated by dividing each time series into one-second windows, with maximal overlap, and then calculating the coefficient of variation of the mean difference between the maximum and minimum amplitudes in each window. A measure of rhythm regularity was calculated by undertaking Fast Fourier Transform to find the distribution of frequencies within each finger tap time series, and then measuring the power of the dominant frequency peak added to the power of the frequencies 0.2 Hz either side of it. A more regular tapping rhythm concentrates power in a narrow frequency band, increasing the power of the dominant frequency peak (and its immediate neighbours), whereas a more irregular tapping rhythm leads to a more widely spread distribution of frequency bands, reducing the power of the dominant frequency peak (and its immediate neighbours).

Spearman correlation coefficients were calculated for the each of the three DeepLabCut measures vs the modal MBRS and MDS-UPDRS clinician ratings. In addition, we calculated the Spearman correlation coefficient of the three computer scores combined (the normalised arithmetic mean of the computer speed score, amplitude variability score and rhythm regularity score) with the MDS-UPDRS rating.

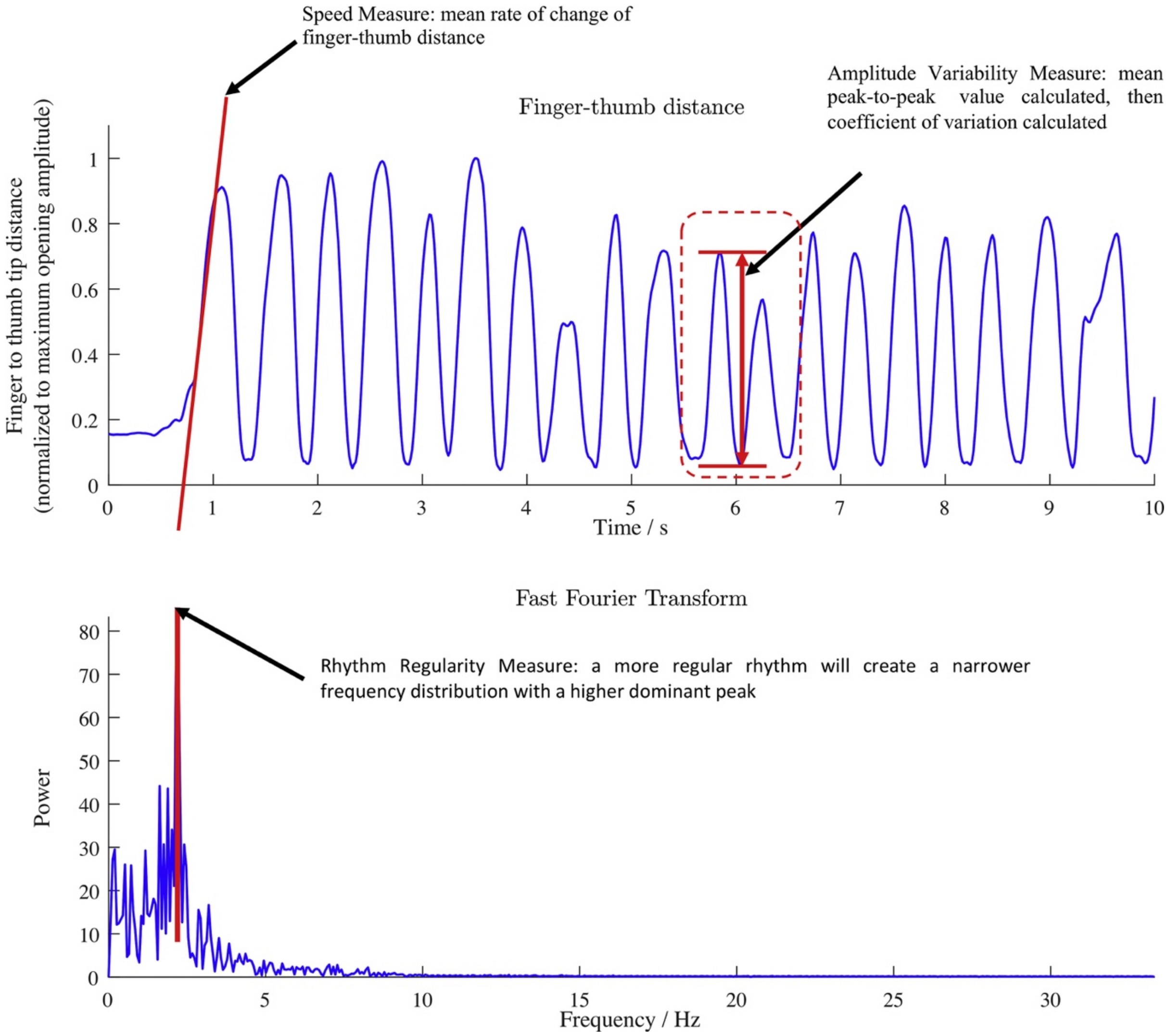


Figure 4.3. Illustration of the three parameters derived from the DeepLabCut finger tip and thumb tip coordinates to measure finger tapping speed, amplitude and rhythm.

Results: Study 1

Characteristics of the participants are presented in **Table 4.1**. MDS-UPDRS-FT scores from 0 to 4 were assigned by two expert clinicians and then categorised into our binary outcome: UPDRS-FT≤1 (no/slight bradykinesia) and UPDRS-FT>1 (mild/moderate/severe bradykinesia). Their assessment matched in 73% of cases (κ = 0.46). **Figure 4.2** shows an example of UPDRS-FT = 0 and UPDRS-FT = 4 for comparison.

	Patients	Controls
Age (Std. Dev.) yrs	67 (10.1)	66 (12.2)
Male/female	26 / 14	12 / 18
Median years since diagnosis	4	-
Median H&Y [IQR]	2 [1, 2.5]	-
H&Y = 1	9	-
H&Y = 1.5	0	-
H&Y = 2	5	-
H&Y = 2.5	1	-
H&Y = 3	4	-
H&Y = 4	1	-
H&Y = 5	0	-
Median UPDRS-FT [IQR]	2 [1, 3]	1 [0, 1]
UPDRS-FT = 0	2	8
UPDRS-FT = 1	11	13
UPDRS-FT = 2	17	7
UPDRS-FT = 3	7	2
UPDRS-FT = 4	3	0

Table 4.1. Study 1 participant characteristics, split by Parkinson's patients and control hands. The modified Hoehn and Yahr (H&Y) is a brief overall clinical rating to describe the stage of symptom progression in Parkinson's (higher number represents more advanced disease) [164]. UPDRS-FT refers to the Unified Parkinson's Disease Rating Scale Item 3.4 (Finger Tapping) [8]. Where raters disagreed the highest of the two UPDRS-FT scores was used. IQR: Interquartile Range.

Two principal components

The performance of each model for the prediction of UPDRS-FT category is shown in **Table 4.2**. The SVM-R model achieved the highest scores in all of our metrics. The other three models perform quite similarly, reflecting the fact that their decision boundaries are close to one another (see **Figure 3**). The test accuracy (estimated using LOO-CV) drops to 0.8 for the SVM-R model, with the other models similarly dropping a few points of accuracy.

Method	Accuracy	AUC	Test Acc	Test Sens	Test Spec
NB	0.74	0.74	0.70	0.67	0.70
LR	0.73	0.73	0.69	0.72	0.65
SVM-L	0.71	0.71	0.71	0.72	0.71
SVM-R	0.84	0.84	**0.80**	0.86	0.74

Table 4.2. Results for each model when predicting whether UPDRS-FT>1 using two principal components. *Accuracy* and *AUC* are estimated from the training 10-fold cross validation and may be considered as upper-bounds. The test accuracy, sensitivity and specificity are estimated using LOO-CV. The emboldened values are the best result for each metric.

In **Figure 4.4** we show each time series plotted in feature-space after dimensionality reduction, marked according to category. We also show the decision boundaries of each method: an unbroken line for NB, dashed for SVM-R, dash-dotted for SVM-L, and dotted for LR.

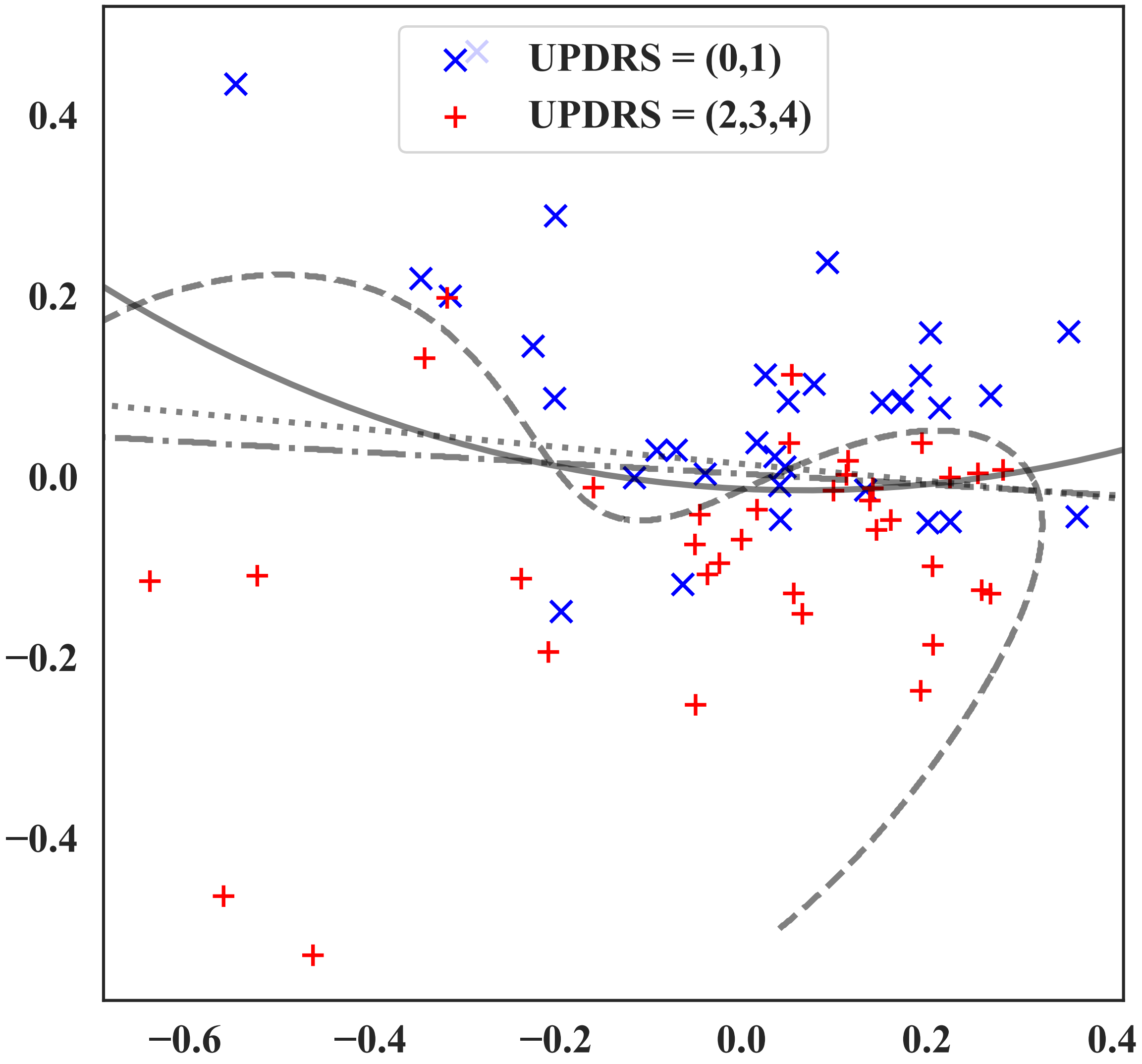


Figure 4.4. Decision boundaries for prediction of UPDRS-FT > 1 using two principal components. The unbroken line is for NB, dashed for SVM-R, dash-dotted for SVM-L, and dotted for LR.

Our second task was the prediction of Parkinson's disease diagnosis itself based upon these features. The performance of each model for this task is shown in **Table 4.3**. Both the NB and SVM-R methods had very similar performance in terms of accuracy and AUC – with NB having better specificity but SVM-R having better sensitivity. Neither LR or SVM-L were competitive for this task unless high sensitivity is desired. A plot of the time series in feature-space, coloured by category, and the decision boundary of each method is displayed in **Figure 4.5**.

Method	Accuracy	AUC	Test Acc	Test Sens	Test Spec
NB	**0.69**	**0.70**	**0.64**	0.58	**0.73**
LR	0.61	0.59	0.61	**0.78**	0.40
SVM-L	0.63	0.60	0.60	**0.78**	0.40
SVM-R	**0.69**	0.68	0.63	0.68	0.57

Table 4.3. Results for each model when predicting Parkinson's diagnosis using two principal components. *Accuracy* and *AUC* are estimated from the training 10-fold cross validation may be considered as upper-bounds. The test accuracy, sensitivity and specificity are estimated using LOO-CV. The emboldened values highlight the best result for each metric.

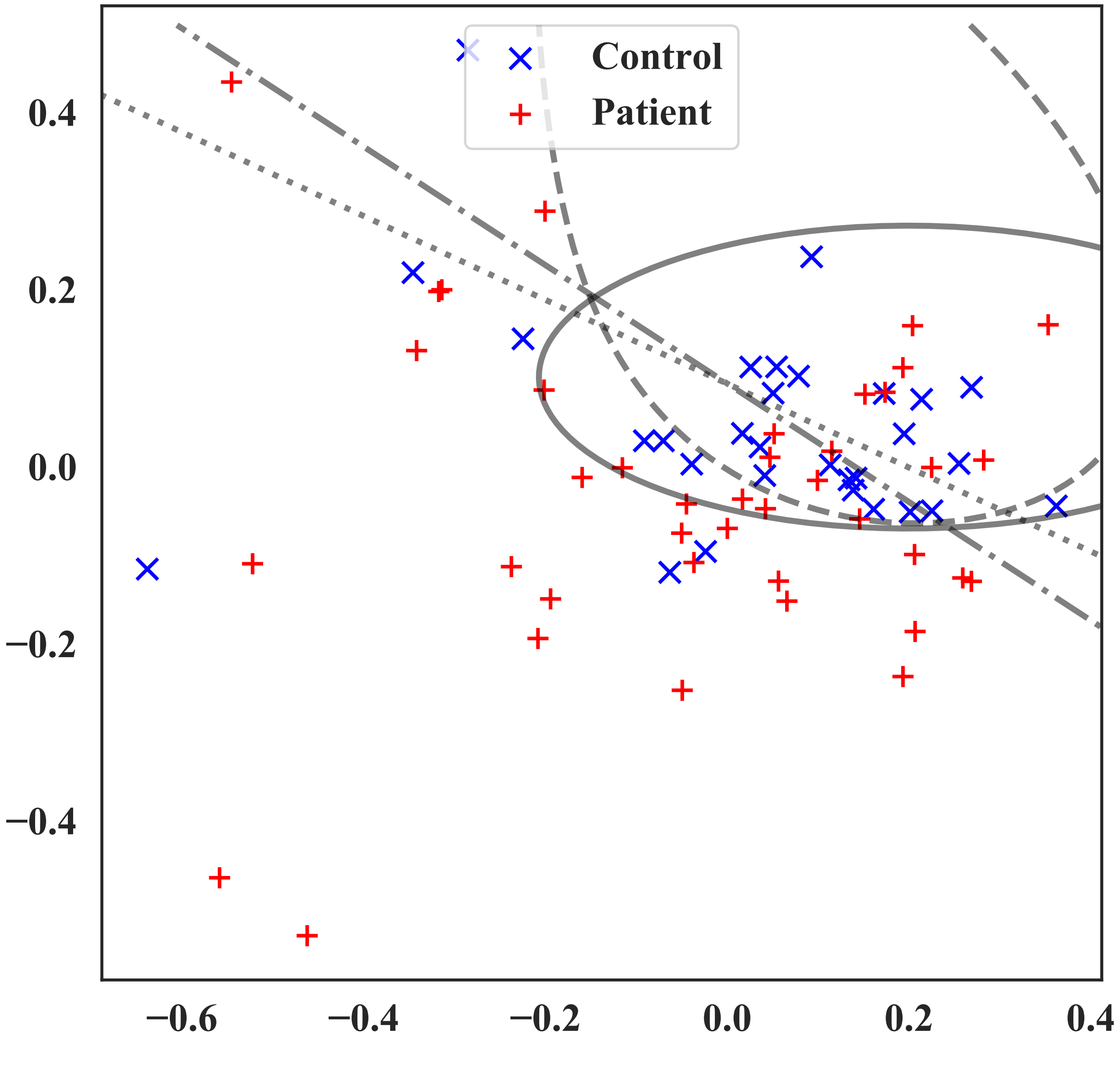


Figure 4.5. Decision boundaries for prediction of Parkinson's diagnosis using two principal components. The unbroken line is for NB, dashed for SVM-R, dash-dotted for SVM-L, and dotted for LR.

Additional principal components

In addition, we experimented with adding additional principal components into the models, training them using the same cross-validation procedure as above.

Tables 4.4 and **4.5** show the resulting accuracy and AUC scores when additional principal components are added, with the test accuracy estimated using LOO-CV.

Method	PCA-2	PCA-3	PCA-4	PCA-5	Test Acc.	P
NB	0.74	0.76	0.79	**0.81**	0.73	0.02
LR	0.73	0.74	**0.76**	0.74	0.73	0.01
SVM-L	0.71	**0.79**	0.76	0.79	0.71	0.01
SVM-R	**0.84**	0.84	0.83	0.81	0.8	0.01

Table 4.4. Resulting accuracy when predicting UPDRS-FT > 1 using 2–5 principal components. The emboldened values highlight the model with highest accuracy in each case The best performing number of principal components was used to estimate the test accuracy with LOO-CV and the permutation test *P*-value.

Method	PCA-2	PCA-3	PCA-4	PCA-5
NB	0.74	0.76	0.79	**0.81**
LR	0.73	0.74	**0.76**	0.74
SVM-L	0.71	**0.79**	0.76	0.79
SVM-R	**0.84**	0.84	0.83	0.81

Table 4.5. Resulting AUC when predicting UPDRS-FT > 1 using 2–5 principal components. The emboldened values highlight the model with highest AUC in each case.

The NB model shows improvement in both accuracy and AUC as more components are included. This did not translate into improved test accuracy, probably indicating model overfitting. The LR and SVM-L models showed minor improvements which again fail to translate into improved test accuracy. The non-monotonic gains in accuracy may be due to the effect of the bias-variance trade-off in this dataset. Overall, the SVM-R model with two principal components had the highest metrics; additional components degrade its accuracy due to the bias-variance trade-off.

The resulting accuracy and AUC when predicting a Parkinson's diagnosis with additional principal components is shown in **Table 4.6** and **Table 4.7**.

Method	PCA-2	PCA-3	PCA-4	PCA-5	Test Acc.	P
NB	0.69	0.63	0.67	**0.74**	0.67	0.46
LR	**0.61**	0.57	0.6	0.57	0.57	0.57
SVM-L	0.63	0.66	0.67	**0.7**	0.63	0.97
SVM-R	0.69	0.8	**0.8**	0.76	0.66	0.49

Table 4.6. Resulting accuracy when predicting Parkinson's diagnosis using 2–5 principal components. The emboldened values highlight the model with highest accuracy in each case. The best performing number of principal components was used to estimate the test accuracy with LOO-CV and the permutation test *P*-value.

Method	PCA-2	PCA-3	PCA-4	PCA-5
NB	0.69	0.64	0.69	**0.74**
LR	**0.59**	0.5	0.57	0.57
SVM-L	0.6	0.65	0.67	**0.68**
SVM-R	0.68	**0.81**	0.79	0.75

Table 4.7. Resulting AUC when predicting Parkinson's diagnosis using 2–5 principal components. The emboldened values highlight the model with highest AUC in each case.

All models except for LR benefited from additional components in terms of in-sample performance but none of these gains translate into improvements in estimated test accuracy. Non-monotonicity in performance as the number of components grows implies there may be some effect from the bias-variance trade-off.

Feature contribution to PCA

Table 4.8 lists the percentage contribution of all of derived features to the first 5 principal components, along with the variance explained by each component.

Feature	SV-1	SV-2	SV-3	SV-4	SV-5
Max peak (Hz)	9.5	14.8	1.3	12.1	9.7
Total ESD	3	16.5	4.5	18.7	19.6
ESD (0-3.06 Hz)	8.3	10.7	23.2	6	4
ESD (3.06-6.12 Hz)	8.3	4	27.8	1.8	8.9
ESD (6.12-9.18 Hz)	12.7	10.8	11.3	8.6	4.6
ESD (9.18-12.24 Hz)	13.2	8	7	10.3	2.9
ESD (12.24-15.3 Hz)	11.8	10.5	4.7	4.1	12.2
ESD (15.3-18.36 Hz)	12	11.6	9.8	9.6	5.9
Maxima-minima ratio	6.4	13	8.6	20.9	20.2
Peak-to-peak std. dev.	14.9	0.1	1.8	8	12.1
Variance explained	37.5	24.1	15.3	7.6	5.5
Cumulative variance	37.5	61.6	76.9	84.5	90

Table 4.8. Contribution of each feature to the first 5 principal components in percentages. The column names SV-n denote contributions to the nth singular vector. ESD is short-hand for Energy Spectral Density.

Our first component explaining 37.5% of the overall variance is comprised primarily of the peak-to-peak standard deviation – measuring variability in rhythm throughout the time series – and the energy spectral density (ESD) in higher frequency bands, which measure jittery movement.

The second component included strong influence from the frequency of the maximal peak (measuring rhythm), the total power in the signal (corresponding to average amplitude across the time series), and the maxima-minima ratio (corresponding to jitter in hand motions).

Misclassified UPDRS-FT categories

We investigated the misclassified examples when predicting UPDRS- FT category using our the SVM-R with two principal components model, to glean insight into where our models may be improved.

This model misclassified 11 examples. 7 were misclassified as mild/moderate/severe bradykinesia (UPDRS-FT > 1) (5 controls, 2 patients). Meanwhile, 4 were misclassified as no/slight bradykinesia (UPDRS-FT 0–1) (1 control, 3 patients). The misclassified examples were close to the decision boundary in 4 cases; for these cases there was expert rater disagreement. All misclassified videos had a UPDRS-FT grade of either 1 or 2, i.e. no large misclassifications occurred.

The time series of two of the examples closest to the decision boundary (one patient and control) are shown in **Figure 4.2**. The two misclassified cases furthest from the decision boundary (one patient and control) are also shown in **Figure 4.2**.

Re-examination of the original videos and optical flow timeseries by two neurologists (SW, JA) identified several potential contributors to this misclassification. First, several videos showed overall hand movement while fingers were held closed between taps, usually a swinging wrist movement preparing for the next tap (and in one case tremor). This created additional small peaks and a more irregular timeseries in videos that showed otherwise regular, smooth finger tapping. Conversely, moving all the fingers 'en masse' tended to create large smooth peaks of optical flow, that reduced the optical flow effect of underlying irregularities in the tapping itself.

Second, a large difference between the speed of finger opening (slower) and closing (quicker) created two distinct optical flow peak sizes/shapes, and a less uniform timeseries, even though the actual tapping was not clearly bradykinetic by UPDRS-FT. Third, our timeseries have a 15s duration, similar to several other objective measures, e.g. [39] but the UPDRS-FT asks raters to judge only the first 10 finger taps. When tapping rate is fast, only a small initial section of the time series is judged by raters, while later tapping changes contribute to the optical flow timeseries.

Finally, a previous study suggested that raters prioritise amplitude and rhythm when judging finger tapping, but pay less attention to speed [27]. With this is mind, we noted that slow but large amplitude movements tended to be classified as UPDRS-FT 0–1 by raters, but UPDRS-FT > 1 by SVM-R, whereas fast but smaller amplitude movements tended to be classified as UPDRS-FT > 1 by raters, but UPDRS-FT 0–1 by SVM-R.

Results: Study 2

The participant and hand video characteristics are given in **Table 4.9**.

	Patients	**Controls**
Age (Std. Dev.) yrs	68 (9.6)	59 (19.4)
Male/female	47/26	22/38
Median years since diagnosis	4	-
Median H&Y [IQR]	2 [1,3]	-
H&Y = 1	32	-
H&Y = 1.5	2	-
H&Y = 2	12	-
H&Y = 2.5	4	-
H&Y = 3	19	-
H&Y = 4	4	-
H&Y = 5	0	-
Median MDS-UPDRS-FT [IQR]	2 [1,3]	1 [0,1]
UPDRS-FT = 0	9	23
UPDRS-FT = 1	20	23
UPDRS-FT = 2	19	12
UPDRS-FT = 3	20	2
UPDRS-FT = 4	5	0
Visible tremor in video	11	0
Visible dystonia in video	2	0

Table 4.9. Study 2 participant (hand video) characteristics. Data is split by Parkinson's hands (n = 73) and control hands (n = 60). H&Y: modified Hoehn and Yahr scale [164]. MDS-UPDRS: Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale, Item 3.4 (Finger Tapping) [8]. IQR: Interquartile Range.

Tracking accuracy

DeepLabCut reliably tracked the finger tapping movements across 133 videoed clinical examinations. **Figure 4.6** shows example frames from a video labelled by DeepLabCut. The mean absolute error of DeepLabCut labelling was 8.39 pixels, i.e. the average distance between manual (human) labels and those predicted by DeepLabCut was 8.39 pixels within a 1920x1080 pixel video frame. **Supplementary Video 4.1** shows examples of four videos labelled by DeepLabCut, including particularly challenging cases with tremor and dystonia also present.

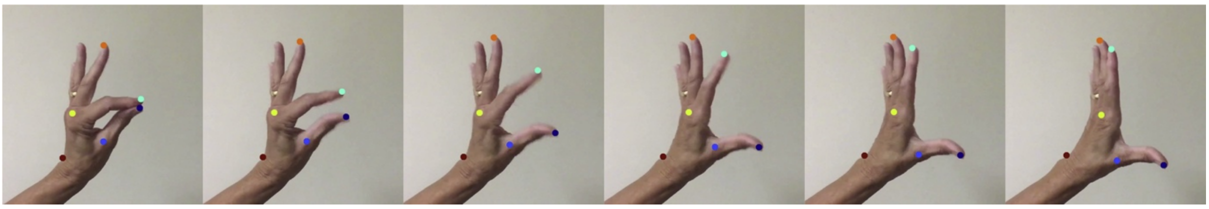


Figure 4.6. Example video frames taken from smartphone video labelled by DeepLabCut, showing the six localization (labelling) points: thumb tip (dark blue), thumb metacarpophalangeal (MCP) joint (light blue), index finger tip (cyan), index finger MCP joint (yellow), middle finger tip (orange), dorsal wrist (red).

Correlation with clinical bradykinesia ratings

The quantitative scores derived from DeepLabCut measurements correlated well with the group neurologists' modal ratings of bradykinesia, **Figure 4.7**. The computer vision measures of tapping speed, amplitude variability and rhythm regularity had good correlations with the respective MBRS clinical ratings for speed (-0.70, p < 0.001), amplitude (0.65, p < 0.001) and rhythm (-0.61, p < 0.001). There were also good correlations between the computer measures of tapping speed (-0.66, p < 0.001), amplitude variability (0.56, p < 0.001) and rhythm regularity (-0.50, p < 0.001) and the MDS-UPDRS clinical rating. The MDS-UPDRS is a composite clinical rating, combining the separate components of bradykinesia, and the three computer measures combined also correlated with MDS-UPDRS (-0.69, p < 0.001).

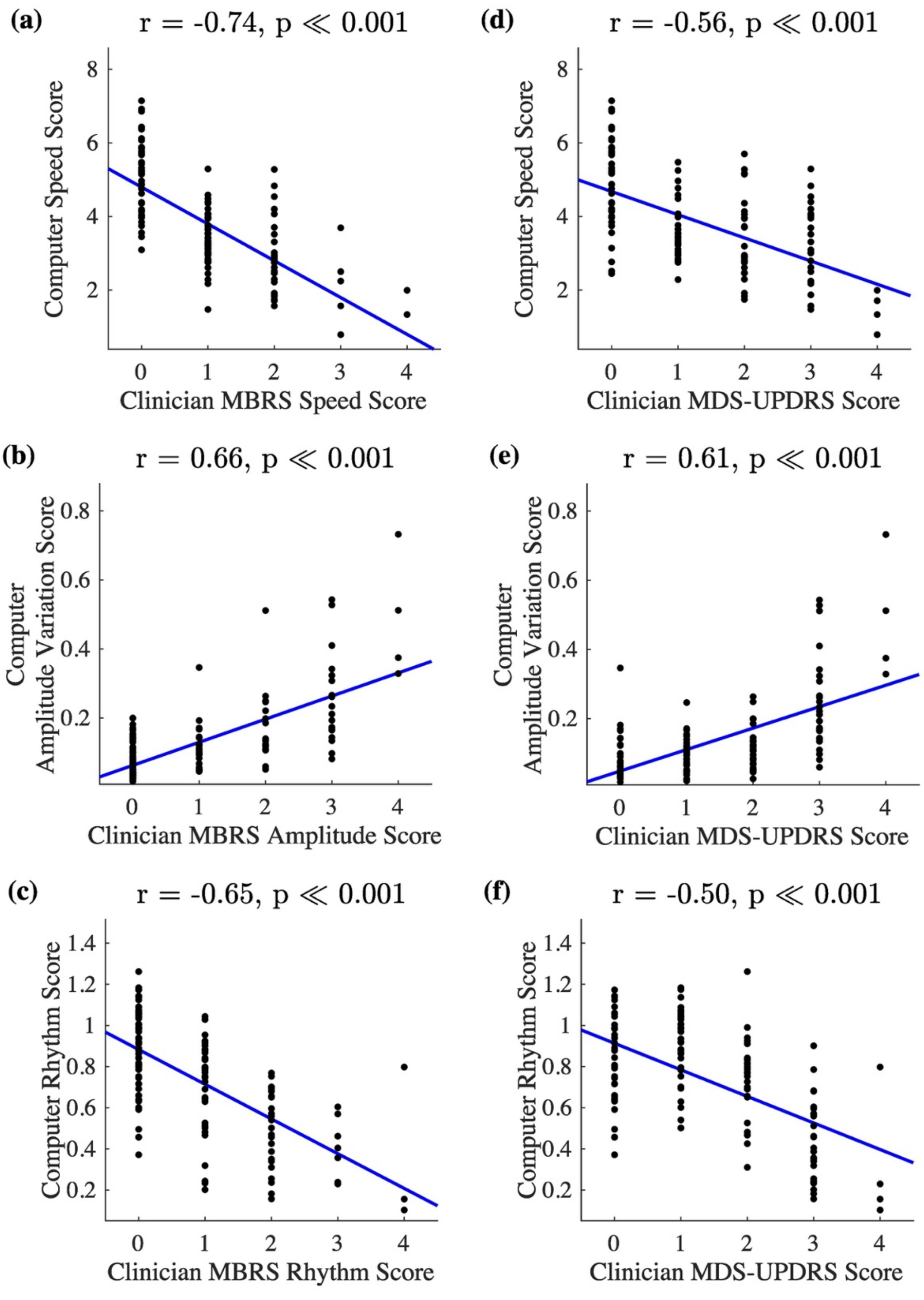


Figure 4.7. Computer measures of finger tapping speed, amplitude variation and rhythm, derived from coordinates produced by DeepLabCut smartphone video tracking, correlate with clinician mode ratings for the same videos. Spearman correlation coefficients shown. MBRS: Modified Bradykinesia Rating Scale. MDS-UPDRS: Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale.

Discussion

This chapter has described two methods that apply computer vision techniques to smartphone video of the finger tapping test for bradykinesia, to derive time-series measurements from the movement of tapping. In both studies, several ‘hand-picked’ features of the time series are used as measures of the theoretical subcomponents of bradykinesia (speed, amplitude, rhythm). These measures are then compared with clinical bradykinesia ratings, and diagnosis. In Study 1, the technique of pixel optical flow was used to measure overall hand movement. Features derived from the resultant time series were combined using principal component analysis and four machine learning techniques were used to classify videos into (a) no or slight bradykinesia versus mild to severe bradykinesia, and (b) participants with Parkinson’s versus control participants. In Study 2, the DeepLabCut neural network was trained to track landmarks on the tapping hand, and these were used to measure relative finger to thumb distance over time. Features derived from this time series were correlated with clinical ratings of bradykinesia.

In Study 1, using 70 videos of finger tapping, the method showed reasonable predictive performance for the presence of mild/moderate/severe bradykinesia (UPDRS-FT > 1). The estimated test accuracy of 0.8 (using an RBF-based support vector machine) is promising in light of the level of agreement between two clinical raters for the same binary judement (κ = 0.46). The method was less successful at predicting the presence of Parkinson’s diagnosis: Naïve Bayes classification obtained an estimated test accuracy of 0.67 using 5 principal components.

In Study 2, using 133 videos of finger tapping, computer vision measures of tapping speed, amplitude variability and rhythm regularity had good correlations with the respective MBRS clinical ratings for speed (-0.70, p < 0.001), amplitude (0.65, p < 0.001) and rhythm (-0.61, p < 0.001). There were also good correlations between the MDS-UPDRS clinical rating and computer measures of tapping speed (-0.66, p < 0.001), amplitude variability (0.56, p < 0.001) and rhythm regularity (-0.50, p < 0.001).

These results support the idea that computer vision can measure finger tapping bradykinesia from standard smartphone video, and that the measures have convergent validity with conventional clinical ratings.

It should be noted that disagreement between these automated methods and clinical experts may be caused when either (i) the clinician is correct and the automated test is wrong, or (ii) the clinician is incorrect and the automated test is right. The results described in **Chapter 3** cast doubt on the ability of human experts to accurately evaluate bradykinesia, so (ii) is highly feasible; such that the reported accuracy may underestimate how well we truly classify or measure bradykinesia. Further improvements in accuracy and generalisability could potentially be achieved in future by using classification algorithms that account for uncertain labels, such as probabilistic SVM [176].

With the worldwide ownership of smartphones so high (e.g. 82% in the UK, 77% in the US) [177] there are no hardware barriers to further use of a computer vision approach. Thus, a major strength of the work is that it has the potential to provide widely available, low-cost bradykinesia detection and / or measurement; without the requirement for new hardware or for patients to directly interact with smartphone apps or computer programs. This is a fundamental difference from most previously published methods that have focussed on wearable sensors measuring finger tapping [27, 38, 44, 47, 178]. In addition, the necessity for sensors to stay fixed on a specific body part has often required the exclusion of tremulous and dyskinetic patients [27, 38, 47, 56, 178]. An automated method broadens access to the measurement of bradykinesia (currently the preserve of a small group of clinicians, principally neurologists). For example, allowing family doctors and nurse practitioners to screen for and monitor the phenomenon has potential resource benefits.

Two particular strengths of Study 2 are the collection of clinical ratings from a large group of movement disorder neurologists, and the use of two rating scales: the MBRS and MDS-UPDRS, such that the clinical ratings were representative and reliable. Clinical rating scales are vulnerable to inter-rater variability but using modal values derived from a large number of blinded clinical raters, and two different gold-standard rating scales, provided a more robust ‘ground truth’ of clinical bradykinesia to evaluate the new test against. As such we were able to demonstrate that clinician and DeepLabCut measures correlate for both individual components of bradykinesia (MBRS) and an overall combination of those components (MDS-UPRDS).

Another advantage of the DeepLabCut approach in Study 2 is that can in principle be applied to any moving body part, in any form of human movement. This means it could be applied to detect and measure other motor features of Parkinson's in a contactless manner such as tremor, gait, or posture. Computer vision in general has potential for application to these signs [179, 180], where wearable methods are limited to one or a few specific clinical signs.

In addition, the 'real life' versatility of a computer vision method in ordinary clinical settings was evaluated in this study because no special efforts were made to optimize participant positioning or filming conditions. The DeepLabCut tracking and measures were robust despite variations in smartphone camera distance, positioning of the hand, skin colour, flexion / extension postures of digits 3 to 5, ambient lighting conditions, and inclusion of a small number of patients with tremor or dystonia during tapping. **Supplementary Video 4.1** illustrates this, with accurate labelling of finger and thumb tip despite superimposed tremor or dystonia. This is particularly encouraging for the relevance of DeepLabCut aiding assessment of Parkinson's in conventional clinical settings.

The studies in this chapter cannot be easily compared with previous literature. Previous studies differ from Study 1 because they have used clinically recognisable features (e.g. tap distance) rather than the overall measure of hand movement that optical flow provides. In contrast, Study 2 used measures with similarity to those in previous studies of 'wearable' devices [27, 38, 44, 47, 178]. However, the results of those previous studies vary widely in terms of strength of correlation or accuracy of discrimination, despite apparently similar methods [26, 40, 181, 182], as described in **Chapter 2.**

At the time of this work, only two previous studies have objectively measured finger tapping using contactless, standard video analysis. One method extracted finger tapping information from only 13 Parkinson's patients, all with advanced disease [149], but it required inclusion of the patient's face in the video, limiting the practical utility. We note that their most predictive feature for UPDRS-FT was a measure of tapping rhythm. This corresponds to Study 1 results, in which the first principal component feature was primarily composed of a rhythm measure (peak to peak variation). Other studies also suggest rhythm measures may be particularly important [27, 43].

A second study involved 60 Parkinson's patients but no controls. It found exceptionally strong correlations between MDS-UPRDS finger tap rating and video measures of tap interval (frequency) (r=0.91), frequency variation (r=0.82) and amplitude (r=-0.94) but not amplitude variation (r=0.39) [135]. The 'ground truth' of the clinical comparison was less robust, based on just two clinical raters, and there were no MBRS ratings; furthermore, some images were blurred in the 24 frame per second videos. A major advantage of our method in comparison is that DeepLabCut is open source software, that can be downloaded and used without the need to write computer code, meaning it is freely available to neurologists with only limited computing knowledge.

A well-studied, non-camera method to measure finger tap bradykinesia is the 'Bradykinesia-Akinesia Incoordination (BRAIN) Test', which involves participants tapping a standard computer keyboard for 30 seconds [12, 183]. This shares advantages with our method in that it involves no special equipment, and could be employed without clinician-patient contact. However, it requires patient motivation to engage with a keyboard tapping task, rather than a simple recording of existing standard clinical examination (and similar requirements for patient motivation apply to patient smartphone apps [46]). Furthermore, 'BRAIN' correlations with MDS-UPDRS finger tapping ratings of 0.44, 0.20, 0.33, 0.03 [183] are weaker than those we report (0.66, 0.56, 0.50, 0.69). One reason for this may be that a computer keyboard cannot record any measure of tap amplitude, and the speed measure is limited to tap frequency, so that some core aspects of bradykinesia cannot be captured by 'BRAIN'.

The relatively poor performance in Study 1 for diagnostic classification of Parkinson's versus controls is to be expected. Indeed, we did not even attempt to discriminate patients from controls in Study 2. In **Chapter 3** we showed that bradykinesia was detected in 24% of control hands when clinical raters are blinded to diagnosis status. When asked to judge patient/control status from the finger tapping video alone, clinician accuracy was 70%, which is very similar to our classification accuracy of 0.67 in Study 1. While bradykinesia is a necessary component of the Parkinson's diagnostic criteria, it is not sufficient in isolation [7]. Clinical Parkinson's diagnosis is by definition a broader assessment than isolated finger tapping performance [7]. In practice, finger tapping bradykinesia is only one of a more comprehensive set of clinical assessments used to diagnose Parkinson's. Consistent for this, in Study 1, for all classifiers for diagnosis, the p-values from the permutation test indicate that similar accuracies may be obtained by chance. While this does not invalidate the result, a much larger training sample would be required to determine whether the classifiers are learning true structure in the data.

The work presented in this chapter has limitations. Clinician MDS-UPDRS ratings are based on 10 finger taps, yet the computer vision time series duration was 15s in Study 1 and 10s in Study 2, similar to some previous technology studies [39]. Some misclassification may have resulted from this difference in assessment time period. Future work could isolate individual tapping epochs [28]. The MBRS ratings are based on 10s of tapping, so they matched time series duration in Study 2.

Like human vision, a simple camera lacks an absolute measure of distance, and our methods cannot capture 3D movement. Rotation of the thumb and finger into a horizontal plane would falsely alter amplitude measurement. However, in practice such movement is rare during tapping, and in Study 2 our normalized measure of finger-thumb distance based on 2D maximum opening distance appeared to capture amplitude, speed and rhythm abnormalities well, given the good correlations with clinical ratings.

Furthermore, even starting with a 2D signal, our approach then discards further spatial and angular information at each frame to produce 1D time series. This has the advantage of reducing the dimensionality of the signal so that real-time processing, even on modest hardware, is practicable. In addition to this, the hand-selection of candidate features was entirely subjective and may have missed important characteristics in the time series. Additional data would allow more sophisticated approaches to automatically learn pertinent features [184].

We have measured amplitude variation in Study 2, but not amplitude decrement (the sequence effect). Previous approaches have measured decrement as a straight line linear regression [47], but decrement would also proportionately affect a measure of amplitude variation, such that the variations in decrement should be captured within our measure.

Furthermore, there are several other previous reports of good correlation between non-decrement amplitude measures using sensors and clinician ratings of finger tap bradykinesia [27].

The continuum of finger tapping performance means that in reality there is a soft boundary between grades, but the use of a binary classifier in Study 1 (e.g. SVM) creates a harder boundary between these classifications, contributing to errors. In future work, we can investigate ‘fuzzy’ or multi-class neural networks to address this.

Relatively few videos were rated as grade 3 or 4 bradykinesia, and it is possible that a greater number of higher grade videos might change the strength of correlations or classification accuracy. Even with accurate tracking, all forms of rating or measuring bradykinesia (including clinician) may potentially be confounded by comorbidities such as joint deformities, pain, dystonia etc. However, with a larger dataset, machine learning techniques are well-suited to separate such additional contributions from underlying bradykinesia.

These studies were exploratory, aimed at proof of concept that bradykinesia related measures could be derived from smartphone video. The protocols were thus relatively simple. The test-retest reliability of this technique has not been addressed in the study but we hope to do so in future. Similarly, we did not measure ‘off’ and ‘on’ medication states (sensitivity to change). Our small sample sizes meant there was insufficient data to test on an independent subset of data. A larger study would allow additional validation tests and greater confidence in the results of machine learning methods.

Chapter 5, Tremor amplification, tremor frequency

Introduction

Tremor can be defined as an involuntary, rhythmic, oscillatory movement of a body part [73]. The Movement Disorder Society (MDS) consensus statement on the classification of tremor uses two main axes [73].

Axis 1 of the MDS classification (**Figure 1.1**) reflects the fact that in most cases, the diagnosis of a tremor disorders remains largely clinical, determined by experienced clinician assessment, rather than a specific investigation. However, clinician diagnosis of tremor is difficult, and several studies show high rates of error. One report asked two movement disorder specialists to distinguish tremor dominant Parkinson's disease from other tremulous movement disorders (atypical tremor and dystonic tremor) using video of clinical examination. There were high false positive (17.4-26.1%) and false negative (6.7-10%) rates for the diagnosis of Parkinson's [92]. Another study reviewed 71 patients previously diagnosed with Essential Tremor, and found that 37% had been misdiagnosed (with the most common true diagnosis being Parkinson's disease) [91].

The application of computer vision techniques to standard video has the potential to augment clinician assessment of tremor in a way that uses widely available hardware, and is contactless, leaving tremor characteristics unaltered by the recording process. This chapter investigates two methods for augmentation of tremor assessment using computer vision applied to video. The first investigates whether amplification of tiny amplitude movement in video might reveal subclinical Parkinson's hand tremor that was not visible in the original video. The second investigates whether the dominant tremor frequency can be measured by tracking the movement of pixels in video of visible hand tremor.

Methods

Video recording

The studies were approved by the London-Fulham Research Ethics Committee of the United Kingdom Health Research Authority, IRAS no. 224848.

Clinical diagnoses had previously been made in routine clinical practice by movement disorder specialist neurologists, according to Movement Disorder Society guidelines [7, 73]. Participants were recruited in order of recent clinic attendance (i.e. there were no special selection criteria). All patients gave written, informed consent for participation.

The participants were seated, and each hand was individually filmed. A smartphone, placed on a tripod, recorded an approximately 60 second video, at 60 frames per second, 1920 x 1080 pixel resolution (standard 'full high definition' smartphone video). Distance from camera to hand was not tightly controlled, but in practice was around 1m, with only the hand and forearm visible within the video frame. A resting tremor recording was made with the forearm on the chair arm and the hand suspended over the end, camera facing the dorsum of the hand, **Figure 5.1**. A postural tremor recording was made with hand and arm extended horizontally forwards from the shoulder, with the camera in a lateral position, **Figure 5.2**. Patients with Parkinson's disease were subjectively and objectively in the 'on' state at the time of video recording.

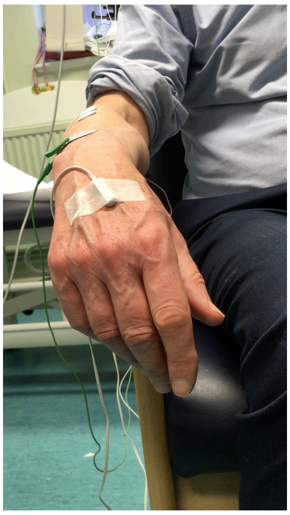


Figure 5.1. An example video frame from a resting tremor recording.

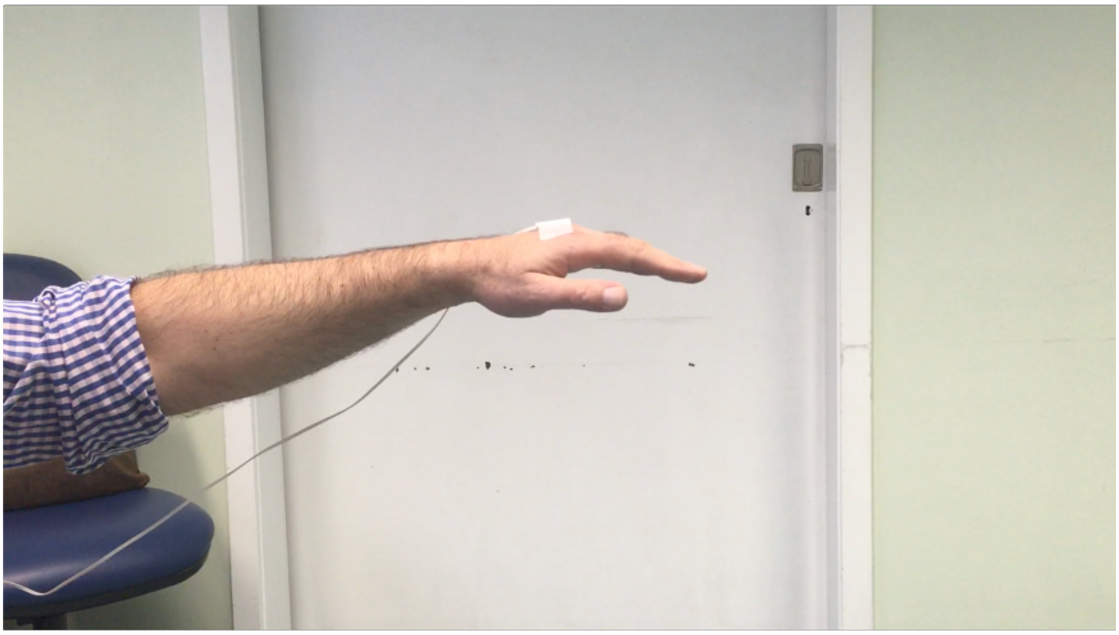


Figure 5.2. An example video frame from a postural tremor recording.

Methods: Study 1 (Eulerian Magnification of hand tremor)

Videos of 48 hands were selected showing hands in a resting position (as described above), in which no tremor was seen at the time of recording. 22 hands were from 11 healthy control participants (no neurological diagnoses or other diagnoses that might cause tremor, no medication likely to cause tremor). 26 hands were from 17 participants with idiopathic Parkinson's disease. The total number of Parkinson's hands was not 34 because 8 participants had obvious visible tremor in one hand, and those hands were excluded from the sample. Each video was 15 seconds duration.

The videos were processed by computer using an Eulerian video magnification algorithm (freely available online) [118], running within MATLAB [171]. This approach can reveal low amplitude movement in video that is difficult or impossible to see with the naked eye. It exaggerates motion by amplifying temporal colour changes at fixed positions. The variation of pixel values over time is amplified in a spatial multiscale manner. The technique first decomposes the video into different spatial frequency bands, and then performs temporal processing on each spatial band. A bandpass filter is used to isolate a temporal frequency band of interest, which we set to between 3 and 7 Hz, to capture the most common frequency range of Parkinson's tremor [73, 75, 76]. The band passed signal is multiplied by a magnification factor, α, which we set to 20. In a final step, the magnified signal is added back to the original to produce a video sequence in which amplified movement is visible.

The authors of the Eulerian magnification technique illustrate the method by using the simple case of a 1D signal, as follows. $I(x,t)$ denotes the image intensity at position x and time t. A video image undergoes translational motion, so that the observed intensities can be expressed with respect to a displacement function $\delta(t)$, so that $I(x,t) = f(x + \delta(t))$. Motion magnification aims to synthesise a signal for an amplification factor α.

$$\hat{I}\ (x,t) = f\ (\ x + (1+\alpha)\ \delta(t)\ )$$

The image at time t, $f\ (x + \delta(t))$ can be written in a first-order Taylor expansion about x, and then let $B(x,t)$ be the result of apply a broadband temporal bandpass filter applied to $I(x,t)$ at every position x:

$$I\ (x,t) \approx f\ (x) +\ \delta(t)\frac{\partial f(x)}{\partial x}$$

and

$$B(x,t) = \delta(t)\frac{\partial f(x)}{\partial x}$$

In the Eulerian magnification process, the bandpass signal is amplified by □□and added back to $I(x,t)$, resulting in the processed signal:

$$\tilde{I}\ (x,t) = I\ (x,t) + \alpha\beta(x,t)$$

When the three equations above are combined, the amplification of the temporally bandpassed signal can be related to motion magnification:

$$\tilde{I}\ (x,t) \approx f\ (x) + (1+\alpha)\delta(t)\frac{\partial f(x)}{\partial x}$$

and

$$\tilde{I}\ (x,t) \approx f\ (x + (1+\alpha)\delta(t))$$

The final equation shows that the processing magnifies motions, because the spatial displacement $\delta(t)$ of the local image *f (x)* at time *t*, has been amplified to a magnitude of $(1 + \alpha)$.

The original videos, and the videos after Eulerian magnification, were rated by three independent UPDRS-certified movement disorder specialist neurologists (clinical raters), in randomised order, with raters blinded to patient / control status. The raters were asked, "does the hand in the video have the appearance of a Parkinsonian tremor ? (yes/no)".

We analysed the proportion of hands correctly classified as Parkinsonian or not before and after amplification by McNemar's test for each rater [185]. McNemar's assesses the dependence of categorical data that are matched or paired. It can be used to test for a change in proportions on a dichotomous trait at two time points in the same participants. On the assumption that an appearance of Parkinsonian tremor is a correct classification for patients but not controls, we assessed correct vs incorrect classification of hand appearance by the rater before and after Eulerian magnification. **Table 5.1** shows the format of McNemar's table for an individual rater.

		Before Amplification	
		Correct classification	Incorrect classification
After amplification	Correct classification	a	b
	Incorrect classification	c	d

Table 5.1. McNemar table for an individual rater asked, 'does the hand in the video have the appearance of a Parkinsonian tremor ? (yes/no)'. Correct classification is defined as Parkinsonian tremor for Parkinson's participant hands and no Parkinsonian tremor for control participant hands.

Chi-square for the proportion of hands correctly and incorrectly classified is given by:

$$\chi^2_{1df} = \frac{(|b - c| - 1)^2}{b + c}$$

The p value is the probability of being wrong if we conclude that Eulerian magnification is associated with a correct classification of hands.

The data from all three clinical raters can be combined and McNemar's test can then be applied as a mixed effects logistic regression model. Our model used correct classification as the outcome, with fixed effects for pre/post-amplification and rater identity, plus a random effect for video number. The fixed effect for amplification provides an odds ratio for the impact amplification has on correct classification estimated over all collected data. The random effect for video number takes into account the range of difficulty in rating (some videos will be easier or more straightforward for the clinician to judge).

Results: Study 1 (Eulerian Magnification of tremor)

Participant details are given in **Table 5.2**.

	Parkinson's	**Controls**
Number of participants	17	11
Hands (no obvious tremor at time of recording) [left:right]	26 [12:14]	22 [11:11]
Mean age [SD], years	69 [11.3]	51 [22.8]
Male:Female	18:8	8:14
Median years since diagnosis	3.8	N/A
No history of previous tremor in filmed hand	8	22
History of previous intermittent tremor in filmed hand	18	0
'Internal tremor' sensation in filmed hand	4	0
Mean age for hands rated 'Parkinsonian tremor' after Eulerian magnification	66	57

Table 5.2. Participant (hand) characteristics. Parkinson's: idiopathic Parkinson's disease. 'No obvious tremor at time of recording' refers to hands in which the investigator (SW) did not see tremor during video recording. SD = standard deviation. 'No history of previous tremor' and 'History of previous intermittent tremor' refer to patient reported symptoms (disease history).

Prior to Eulerian magnification, the original videos were judged to show Parkinsonian tremor in the following number of Parkinson's participant hands: 0/26 (rater 1); 1/26 (rater 2); 0/26 (rater 3), and the following number of control participant hands: 0/22 (rater 1); 1/22 (rater 2); 1/22 (rater 3). The single control hand rated as tremulous by rater 2 was also the single hand rated as tremulous by rater 3.

After Eulerian magnification, the following number of Parkinson's hands were judged to show Parkinsonian tremor: 14/26 (rater 1), 6/26 (rater 2), 7/26 (rater 3), and the following number of control participant hands were judged to show Parkinsonian tremor: 7/22 (rater 1); 4/22 (rater 2); 3/22 (rater 3). **Supplementary Video 5.1** shows an example of a Parkinson's participant hand and a control participant hand before and after Eulerian magnification. Each participant hand appears atremulous in the original video, but only the Parkinson's participant hand appears to have a typical Parkinsonian tremor after movement amplification.

The group mixed effects model combining scores for all three raters showed a significantly higher proportion of correctly classified hands after Eulerian magnification, (OR = 2.67; CI = [1.39, 5.17]; p < 0.003), **Figure 5.3**. For each of the three individual raters, the proportion of correctly classified hands increased after Eulerian magnification, although did not reach significance when analysed in isolation (p = 0.08, 0.36, 0.09).

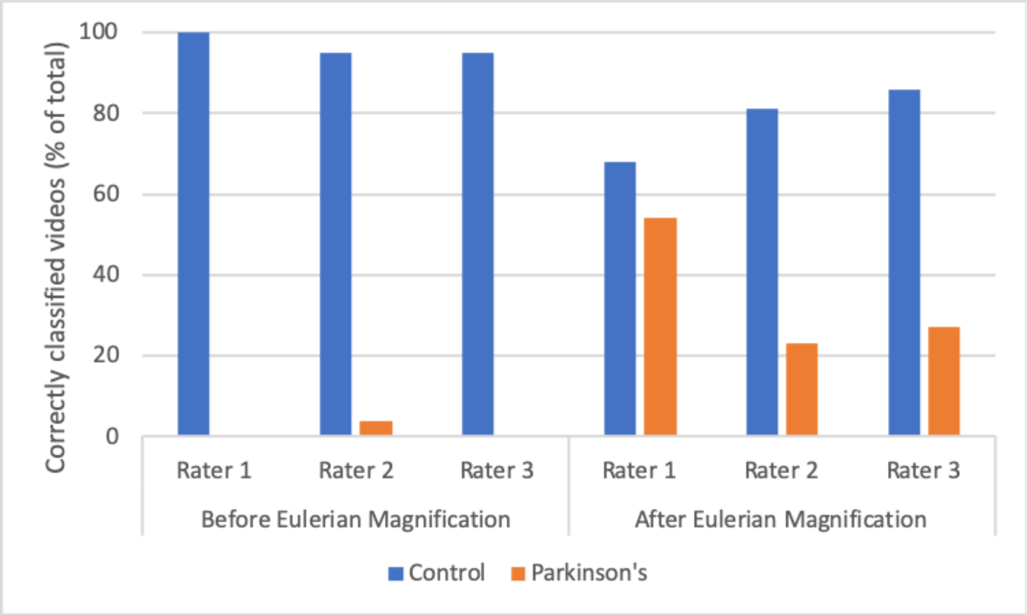


Figure 5.3. Eulerian video magnification improves clinician classification of hands as Parkinsonian or control**.** The overall proportion of correctly classified hand videos is increased after Eulerian magnification, p<0.003 (McNemar test mixed effects logistic regression model). Correct classification is defined as Parkinsonian tremor in Parkinson's hands and no Parkinsonian tremor in control hands (raters blinded to diagnosis). Note that some control hands were incorrectly classified after amplification (lower blue bars on right side of graph), but *all* Parkinson's hands except one (for one rater) were incorrectly classified prior to amplification (lack of orange bars on left side of figure).

Methods: Study 2 (Tremor Frequency)

A convenience sample of 16 patients from Leeds Teaching Hospitals NHS Trust (United Kingdom) participated in the study, with the following established diagnoses: 5 essential tremor, 2 functional tremor, 9 Parkinson's disease.

Only those videos with visible tremor were used, making a total of 40 videos.

In accordance with standard methods of tremor accelerometry [86], a single axis accelerometer ('Natus neurology tremor sensor') was attached to the dorsum of each participant's hand, aligned with the long axis of the hand. Acceleration was recorded contemporaneously with the video (without an exact time-lock mechanism), at a sample rate of 3.84 kHz.

The first 2 seconds of the video and accelerometer recordings were removed to reduce any voluntary movement artefacts as the patient settled into position. The endings of both recordings were cropped at 60 seconds.

The videos were processed with custom-written MATLAB code [171] [https://github.com/DrStefanWilliams/tremor-optical-flow]. It allows a bounding box to be manually drawn around the hand region for one frame of each video, together with a line to mark the long axis of the hand, a process that takes approximately 15 seconds for the user to complete.

Within the bounding box drawn on the video, hand movement was quantified using an optical flow method. Optical flow can be defined as "the distribution of apparent velocities of movement of brightness patterns in an image" [108]. Relative motion of objects and the viewer creates optical flow, as outlined in Chapter 1. In the hand tremor videos, the camera is static, and so the participant's tremulous hand creates a moving brightness pattern. Optical flow algorithms relate the change in image brightness at a point to the motion of the brightness pattern. Horn and Schunck's 1981 method [108] is based on the following constraints. If brightness at a point (*x, y*) at time *t* is denoted by *E(x, y, t)*, then the unknown optical flow velocity can be represented as *u* and *v*:

$$u = \frac{\mathrm{d}x}{\mathrm{d}t} \quad \text{and} \quad v = \frac{\mathrm{d}y}{\mathrm{d}t}$$

The brightness can be related to *u* and *v*:

$$E_x u + E_y v + E_t = 0$$

Which can be written as:

$$\left(E_x, E_y\right) \cdot (u, v) = -E_t$$

So that the component of movement in the direction of the brightness gradient equals:

$$\frac{E_t}{\sqrt{E_x^2 + E_y^2}}$$

Most commonly, video involves opaque objects of finite size undergoing rigid motion or deformation (which is the case for a tremulous hand). This means neighbouring points on an object have similar velocities, so that the velocity of brightness patterns in the image varies smoothly almost everywhere. An additional smoothness constraint can be introduced, that minimises the square of the magnitude of the gradient of the optical flow velocity:

$$\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \quad \text{and} \quad \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2$$

To create a one-dimensional time series from the optical flow results, we used a Histograms of Optical Flow method [109]. In this technique, once optical flow is computed for every frame of the video, each flow vector is binned according to its primary angle from the horizontal axis and weighted according to its magnitude, **Figures 1.3 and 1.4**. All optical flow vectors *v = [x,y]*$^{\mathrm{T}}$, with direction

$$\theta = \tan^{-1}\left(\frac{y}{x}\right)$$

in the range

$$-\frac{\pi}{2}+\pi\frac{b-1}{B}\leq\theta<-\frac{\pi}{2}+\pi\frac{b}{B}$$

will contribute by

$$\sqrt{x^2+y^2}$$

to sum in bin *b,* 1 ≤ *b* ≤ *B*, out of total of *B* bins [109].

We defined two bins in our histogram of optical flow for the tremor videos, with one bin on either side of a line marking the long axis of the hand. This provided a time series for hand movement in two directions, perpendicular to the long axis of the hand. We converted this into a one-dimensional time series by subtracting the movement in one direction from that in the opposite direction, for each video frame (Figure 5.4(B), left hand graph).

The resultant accelerometer and video time series were converted into the frequency domain using a Fast Fourier Transform (FFT)[4]The Discrete Fourier Transform (DFT) is given by the following equation:

$$Sx(f)=\int_{-\infty}^{+\infty}x(t)e^{-j2\pi ft}dt$$

S_x(f) is the output of the Fourier Transform in the frequency domain

x(t) is the input time domain function

2πf is the frequency in radians per second

The Fast Fourier Transform calculates the DFT in a more computationally efficient way. Where the DFT takes O(n2) time, the FFT takes O($n$log$n$) steps. An FFT re-expresses the DFT of an arbitrary composite size N = N1N2 in terms of N1 smaller DFTs of sizes N2, recursively. The Cooley-Tukey FFT algorithm divides the DFT into two smaller parts as follows:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i\,2\pi\,k\,n/N}$$

$$= \sum\nolimits_{m=0}^{N/2-1} x_{2m} \cdot e^{-i\,2\pi\,k\,(2m)/N} + \sum\nolimits_{m=0}^{N/2-1} x_{2m+1} \cdot e^{-i\,2\pi\,k\,(2m+1)/N}$$

$$= \sum\nolimits_{m=0}^{N/2-1} x_{2m} \cdot e^{-i\,2\pi\,k\,m\,/(N/2)} + e^{-i\,2\pi\,k/N} \sum\nolimits_{m=0}^{N/2-1} x_{2m+1} \cdot e^{-i\,2\pi\,k\,m/(N/2)}$$

Python ('scipy.fftpack') [161] was used to perform the Fast Fourier Transform (FFT) for accelerometer and video time series. We then removed frequencies below 2 Hz (low-frequency drifting movements) and above 14 Hz (higher than hand tremor) from the analysis. Examples of accelerometer and video signal time series and their FFT are shown in **Figure 5.4**.

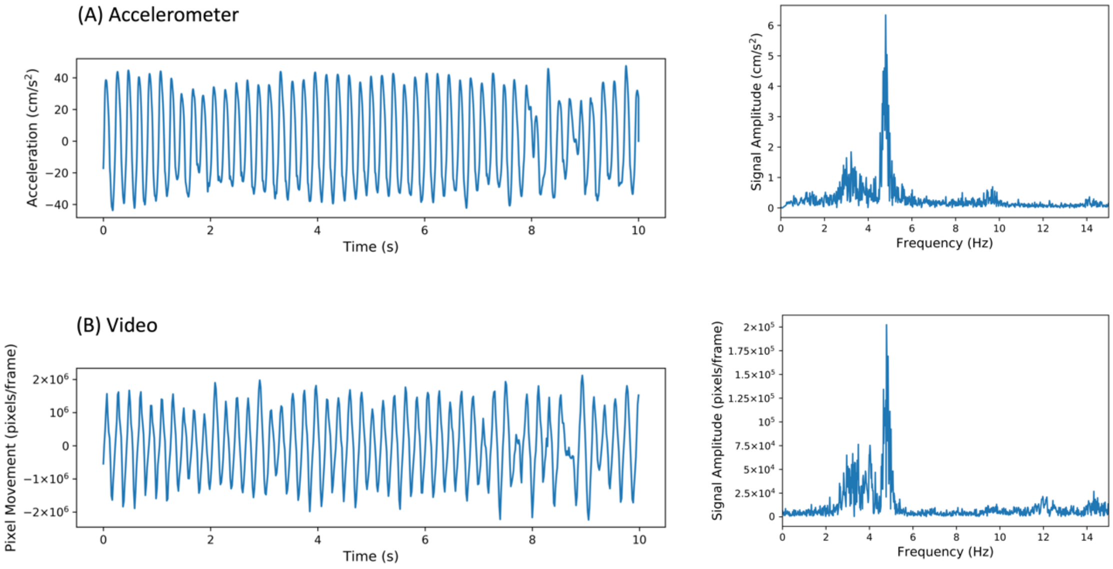


Figure 5.4. Examples of time series and frequency domain after fast Fourier transform, for **(A)** accelerometer (dominant frequency 4.79 Hz) and **(B)** video (dominant frequency 4.80 Hz). (Video signal derived from pixel movement via Histograms of Optical Flow method.)

The recordings from one patient with functional tremor showed widely dispersed frequency distributions (on both accelerometer and video derived FFT) for the 60 second recording, without any obvious dominant peaks. This is consistent with the known clinical variability of functional tremor over time (i.e. the frequency can vary greatly over 60 seconds, so that there is no single, consistent, dominant frequency). These 3 recordings were removed from the final analysis.

For the remaining 37 videos, the dominant frequency derived from video was compared with the dominant frequency derived from the accelerometer. The mean absolute error was calculated (i.e. the mean absolute difference between video and accelerometer frequencies, without the direction of individual differences). We then undertook a Bland-Altman analysis [130]. The Bland-Altman method was originally developed because of the recognition that a correlation coefficient is not a measure of agreement (correlation measures the strength of relationship between two variables, but not the agreement between them). Measurements from two different devices can be perfectly correlated, without the measurements agreeing. Furthermore, correlation strength increases if the range of samples is broader (without any increase in agreement) [130].

The Bland-Altman plot is derived from the following calculations: the bias (mean difference) between video and accelerometer, d; the standard deviation of the differences, *s*; and the 95% confidence intervals for the mean difference, which are referred to as 'limits of agreement', and calculated as d -1.96*s* for the lower limit of agreement and d +1.96*s* for the upper limit of agreement. In addition, the 95% confidence intervals (CI) for each of the limits of agreement can be calculated as follows [128]:

CI for mean – 2SD = $(d - 2\mathrm{SD}) \pm t \times (\sqrt{3\mathrm{SD}^2/n}$

CI for mean + 2SD = $(d + 2\mathrm{SD}) \pm t \times (\sqrt{3\mathrm{SD}^2/n}$

In a Bland-Altman analysis, it is considered ideal to decide a clinically acceptable size range for the 95% limits of agreement, prior to the analysis, i.e. how far apart measurements can be without causing significant clinical difficulties [128]. We considered ±0.5 Hz to represent *a priori* clinically acceptable 95% limits of agreement for tremor.

A minority of recordings showed two distinct and prominent frequency peaks. We considered a frequency spectrum to have two peaks if a secondary peak was at least 70% of the dominant peak amplitude, and if the frequency of the secondary peak was at least 0.5 Hz from the primary peak, see example in **Figure 5.5**. Given that clinical assessment of tremor assumes only one dominant frequency, we selected the peak with highest frequency in these scenarios.

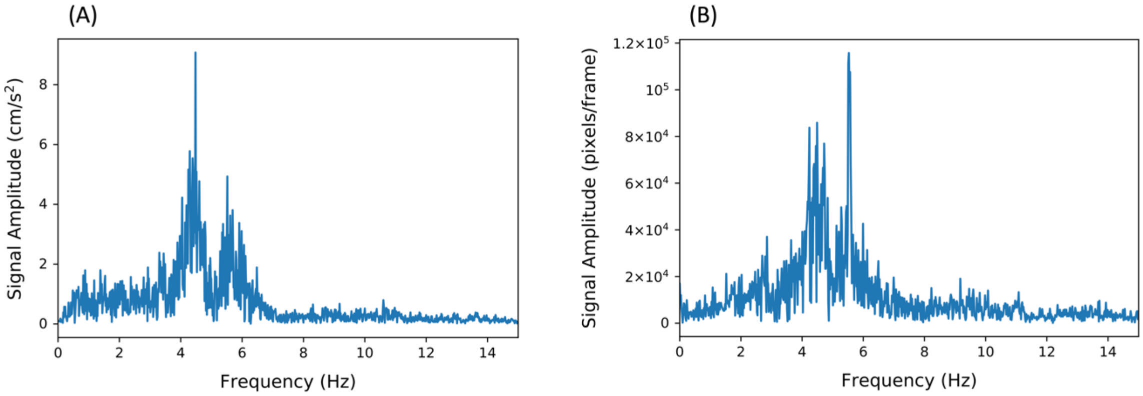


Figure 5.5. Example of frequency distributions (Fast Fourier Transform) with two peaks. The two peaks show the same frequencies in accelerometer and video, but each has a slightly higher power than the other in the alternative methods of measurement. **(A)** Accelerometer shows a first dominant peak of 4.47 Hz and a second dominant peak of 5.51 Hz; **(B)** video shows a first dominant peak of 5.53 Hz and a second dominant peak of 4.49 Hz. The distributions and dominant frequencies are very similar. Where a second dominant peak was >70% the size of the first dominant peak (and separated by >0.5 Hz), we compared the highest frequency peak of the two in both methods of measurement (in this case 5.51 Hz and 5.53 Hz).

To test for an inverse correlation between tremor amplitude and the accuracy of tremor frequency derived from video, we undertook two approximate measures of tremor amplitude. The first calculated an approximate measure of movement amplitude from the accelerometer signal. After converting the units from microvolt to cm/s^2, we used Python [161] to scale the accelerometer data to have a mean of 0, and then used the midpoint rule to estimate the first integral (velocity), followed by eliminating drift by subtracting the line of best fit (using the ordinary least squares method). We then used the midpoint rule again to estimate the second integral, followed by calculating the standard deviation of mean displacement from the baseline. As a second measure of tremor amplitude, two neurologists (SW, JA) clinically rated postural and rest tremor amplitude from the videos, according to MDS-UPDRS items 3.15 and 3.17 respectively (grade 0 = no tremor, grade 1 = <1cm amplitude, grade 2 = 1-3cm amplitude, grade 3 = 3-10cm amplitude, grade 4 = >10cm amplitude). The median amplitude rating was MDS-UPDRS grade 2 (interquartile range 1-3). Spearman's correlation coefficient was calculated to test for relationships between these amplitude measures and video accuracy (absolute error, Hz).

Results: Study 2 (Tremor Frequency)

Participant and video details are given in **Table 5.3**.

	Essential tremor	**Parkinson's disease**	**Functional tremor**
Number of participants	5	9	1
Age (Std. Dev.) years	63 (10)	66 (12)	41
M:F	1:4	7:2	0:1
Resting tremor hand recordings	9	14	2
Postural tremor hand recordings	7	5	0
Median dominant accelerometer frequency (IQR), Hz	5.5 (5.1-5.9)	5.4 (4.1-5.9)	8.0

Table 3. Summary of Participant Characteristics (IQR = Interquartile Range)

The dominant tremor frequencies from video (pixel optical flow) showed a mean absolute error of 0.10 Hz (standard deviation ±0.16 Hz) compared with the accelerometer frequencies. In 36 out of 37 videos (97%) there was less than 0.5 Hz difference between the computer vision and accelerometer frequency measurements. Bland-Altman analysis of the dominant frequencies from video vs accelerometer showed a mean difference (bias) of -0.01 Hz, **Figure 5.6**. with 95% limits of agreement -0.38 Hz to 0.35 Hz. The 95% confidence intervals for the limits of agreement were -0.48 Hz to -0.31 Hz for the lower limit and 0.28 Hz to 0.46 Hz for the upper limit.

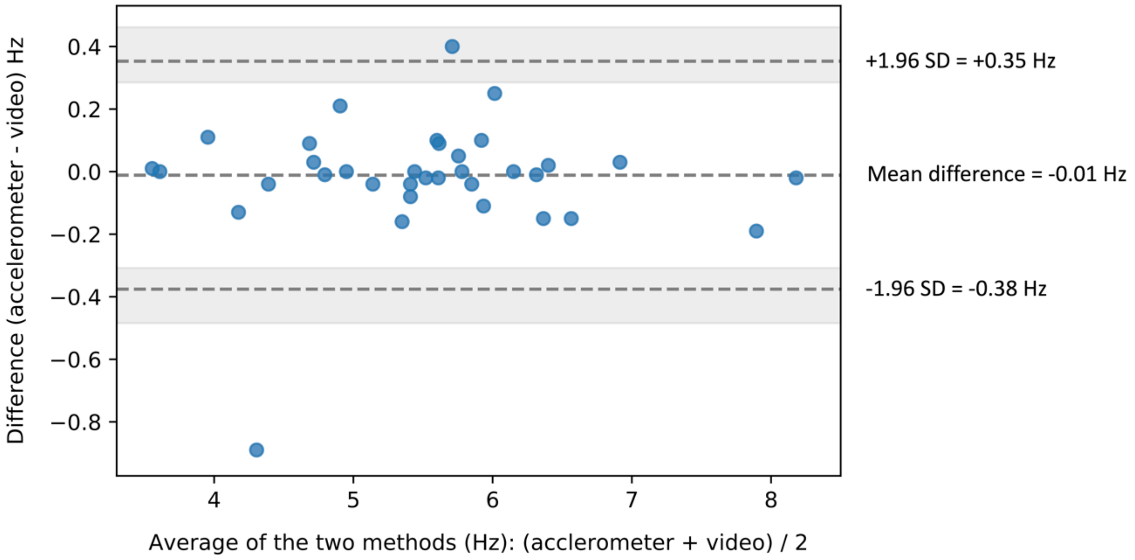


Figure 5.6. Bland-Altman Plot showing the agreement between tremor measurements derived from video (pixel optical flow) and accelerometer. Dashed horizontal lines show mean difference (bias) and the 95% limits of agreement. Grey bands show the 95% confidence intervals for the limits of agreement. The relative outlier (0.89 Hz difference) was a Parkinson's hand at rest.

15 of the 37 videos were rated as <1cm amplitude by both clinician raters (i.e. MDS-UPDRS grade 1 or 0). No significant inverse correlation was found between the absolute error of video tremor frequency (Hz) and tremor amplitude, either measured by clinician amplitude rating (Rater 1: Spearman's rho -0.07, p=0.70; Rater 2: Spearman's rho -0.15, p=0.37) or accelerometer displacement (Spearman's rho 0.25, p=0.14), **Figure 5.7**. **Supplementary Video 5.2** shows an example of two tremors with low amplitude (UPDRS grade 1) in which there was excellent agreement between the video and accelerometer frequencies (6.41 vs 6.39 Hz for the first hand shown in the video, and 5.60 vs 5.62 for the second hand shown in the video).

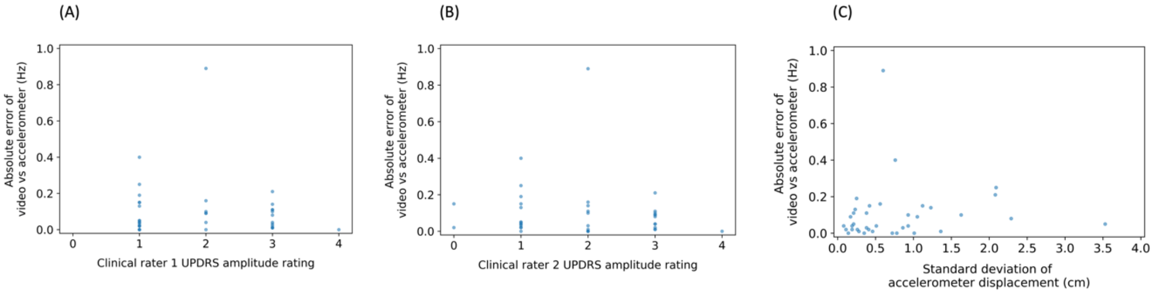


Figure 5.7. No significant inverse correlations were found between the absolute error of video tremor frequency and measures of tremor amplitude. Y axes show absolute error of video tremor frequency (the difference between video and accelerometer tremor frequency, Hz). X axes on graphs **(A)** and **(B)** show the grade of tremor amplitude rated by two neurologists ('rater 1' and 'rater 2'), using MDS-UPDRS items 3.15 for postural tremor and 3.17 for rest tremor (grade 0 = no tremor, grade 1 = <1cm amplitude, grade 2 = 1-3cm amplitude, grade 3 = 3-10cm amplitude, grade 4 = >10cm amplitude). The x axis on graph **(C)** shows the standard deviation of mean accelerometer displacement from baseline. Spearman's rho and p values for each graph: (A) rho -0.07, p=0.70; (B) rho -0.15, p=0.37; (C) rho 0.25, p=0.14.

Discussion

This chapter describes two main findings. The first is that amplification of video pixel movement (Eulerian magnification) reveals apparent subclinical tremor of both Parkinson's hands and control hands that appear still in the original (pre-amplification) video. However, a greater number of Parkinson's hands appear tremulous after amplification, and the proportion of hands that clinicians correctly classify as Parkinsonian is increased after amplification (p<0.003, McNemar mixed effects logistic regression model). The second finding is that the movement of pixels in video, determined by an optical flow algorithm, can be used to measure the dominant frequency of hand tremor, and this shows excellent agreement with a gold standard measure of tremor frequency, accelerometery (95% limits of agreement -0.38 Hz to +0.35 Hz).

Eulerian Magnification Study

Tremor can be defined as an involuntary, rhythmic, oscillatory movement of a body part [73]. Several mechanisms can underlie such movement, either at a subclinical amplitude that is too small to be seen, or a larger, visible clinical amplitude [86, 186]. Firstly, all objects, including hands, will oscillate at a particular frequency when receiving energy and not restrained. This is natural or resonant frequency, and 'mechanical' tremor. In pure form, it is subclinical and invisible. The specific frequency depends on the object's inertia and stiffness, whilst the energy that drives it in hands is derived from the ballistic cardiac impulse and irregularities in the firing rate of motor units [86, 186]. Secondly, factors that increase the gain of the monosynaptic stretch reflex (e.g. fatigue, stress, adrenergic medication) can create an exaggerated reflex response that is both triggered by mechanical tremor and in turn exacerbates mechanical tremor. This can create a clinical, visible tremor, termed the 'mechanical-reflex' component [86]. Thirdly, structures in the central nervous system can undergo abnormal oscillatory activity that is transmitted along the motor system, creating a tremor of 'central' origin [86, 187].

Physiological tremor usually has only a mechanical component, and in the hand the frequency typically ranges between 6 and 12 Hz. Enhanced physiological tremor can contain mechanical, mechanical-reflex and central components with a broad range of frequencies. Parkinson's tremor is mainly central, with a frequency often quoted as 4 to 7 Hz.

The results of Eulerian magnification applied to video of resting hands can be interpreted in terms of these different tremor mechanisms. All hands undergo a degree of subclinical, low-amplitude mechanical tremor, and there is no reason to expect a systematic difference in this between Parkinson's and control hands. Since the frequency amplified was 3 to 7 Hz, and mechanical tremor is described as 6 to 12 Hz, then one explanation for the apparent revealed tremor in both patient and control hands is that subclinical physiological (mechanical) tremor is being amplified in both groups. The clinicians viewing the videos saw 'parkinsonian' tremor because the hands were resting and the band of amplification was chosen at the typical frequency range of a Parkinsonian tremor. However, the amplification of mechanical tremor does not explain why post-amplification Parkinsonian tremor was seen in a greater number of Parkinson's than control hands, and an increased proportion of hands were correctly classified as Parkinsonian. One possible explanation for this is that some of the Parkinson's hands had an additional subclinical central (Parkinson's) tremor component, that was also amplified. Thus, an interpretation of the findings is that Eulerian magnification reveals physiological tremor in control hands but reveals both physiological and subclinical Parkinson's tremor in Parkinson's hands. This would suggest that subclinical pathological tremor is detected at a group level. As such, Eulerian magnification may represent a first step towards contactless visualisation of subclinical pathological tremor.

There are significant limitations to the results of Eulerian magnification. Firstly, even if the results reflect the detection of subclinical pathological tremor in Parkinson's hands at a group level, there is no obvious clinical utility to detecting this group difference. In its current form, Eulerian magnification is not a diagnostic test or biomarker (it lacks specificity and sensitivity). The study tested 'classification' of Parkinson's and control hands, but the implications of misclassification are not the same for control and Parkinson's hands. Erroneously classifying healthy control hands as Parkinsonian runs the risk of considerable psychological distress, so that any future development of Eulerian magnification would require better separation of signal from 'noise', to reduce false positive appearances. Secondly, the Parkinson's patient group was heterogenous, e.g. the presence/absence of 'internal tremor' sensation, or visible tremor in the contralateral hand, and it may be that more uniform subgroups could identify a stronger or more specific effect of Eulerian magnification. Related to this, some idiopathic Parkinson's patients never develop clinical tremor, and there is a recognised distinction between the predominant resting tremor subtype and the postural instability gait subtype of the disease [78]. Thus, the assumption that parkinsonian tremor after Eulerian magnification is the correct classification for Parkinson's hands may not be an appropriate one for all patients. The algorithm cannot be expected to reveal subclinical movement that was never there and may never be so.

Thirdly, we did not record accelerometer or EMG as a 'gold standard' comparator. However, previous work on subclinical pathological tremor has not used these methods, so their usefulness or 'gold standard' status in this regard is unclear. Accelerometer has been used to study subclinical postural hand tremor in normal (control) participants. The findings showed that the frequency spectrum patterns of the tremor varied greatly among the participants, with some showing single finely tuned peaks, and others showing multiple broad peaks, with two thirds showing complex patterns [188]. Upper and lower limits of the tremor frequency power also varied greatly. This suggests that the Eulerian magnification of control hand videos would be expected to produce varied results within the group, complicating any attempt to separate control signal from pathological Parkinson's signal. It is consistent with the fact that apparent parkinsonian tremor was seen in some but not all controls after amplification in our study. Furthermore, the peak tremor frequency among normal participants had a mean around 7 Hz below age 70 and 6 Hz above age 70. These are within the typical Parkinson's range of tremor and within the band of tremor amplified in our study (3-7 Hz). Indeed, the range of Parkinson's rest tremor frequency has a broad and relatively vague definition, variously quoted at 3-6 Hz, 4-6 Hz, and 4-7 Hz [73, 75, 76]. A technique that selects a frequency band will inevitably amplify both subclinical physiological tremor as well as any subclinical Parkinson's tremor.

Subclinical tremor in Parkinson's has been studied with a displacement laser transducer, attached to the tip of the index finger, and compared with recordings from matched control participants [189]. A series of characteristics were measured from the raw displacement data and an analysis of variance (ANOVA) was conducted to test for significant patient vs control group differences. Subclinical resting tremor in Parkinson's was found to differ from control physiological tremor in terms of amplitude fluctuation, frequency dispersion, harmonic index, and proportional power in 4 to 6 Hz [189]. However, these are group differences, further reinforcing our findings that suggest it is difficult to separate pathological from physiological tremor at an individual level. Furthermore, it differs from the work in this chapter because it does not involve signal amplification. It would be more analogous to looking for numerical differences in the original pixel movement rather than amplifying the movement for clinician visual judgement. This leads to a potential fundamental problem with Eulerian magnification, which is that it may simply be mainly amplifying 'noise'. The frequency distributions of visible tremor in the second half of this chapter show some signal at most frequencies: the baseline is rarely zero, e.g. **Figure 5.4**, **Figure 5.5**. This 'noisy' baseline is also seen in the frequency distribution of physiological tremor [188]. We amplified movement by a factor of 20 within a frequency band, and it is possible that the resulting video appearances could be largely due to baseline noise dominating the visible movement when it is amplified to that degree.

The mechanical and central components of postural hand tremor can be separated by weight-loading, whereby the attachment of weight to the hand decreases the frequency of the mechanical component of a tremor, but not the central component [86]. A future protocol could involve amplification of hands in the postural position and comparing the appearance with and without weight loading to try to separate out central from mechanical components within the amplified signal.

Finally, the rate of correct 'classification' by clinicians as parkinsonian or not is affected by the relative number of Parkinson's and control participants. Prior to amplification, all Parkinson's patients will be classified incorrectly and all controls correctly, so that any improvement with amplification will depend on the relative proportions of Parkinson/control hands at baseline (more controls sets a higher baseline accuracy of classification from which to improve). There were slightly fewer control hands in our study. However, in practice this is unlikely to have played a major role, as the numbers were almost equal and the proportional increase in parkinsonian tremor appearance was greater within the Parkinson's group for all raters.

Tremor Frequency Study

The second study in this chapter demonstrated that visible hand tremor frequency can be measured from video with good agreement between video and the gold standard measure of accelerometer. Frequency is one of the clinical features within Axis 1 of the Movement Disorder Society consensus statement on the classification of tremor, **Figure 1.1** [73], and measurement of frequency can aid the diagnosis of tremor disorder [86]. However, it is not possible to accurately estimate tremor frequency by eye during clinical examination. Unlike accelerometers (clinical or within smartphones), our video method is contactless, does not involve specialist equipment or patient interaction with an app, and simply visually records what the clinician sees during a standard neurological examination. The hardware required already exists in the pocket of most clinicians, so that our approach is an equitable one, that bypasses many of the usual cost and geographical barriers. In essence, our method is largely automated, without a need to tightly control camera distance, lighting or background. There are only three manual procedures that are currently inherent to our method: (1) setting the smartphone camera to record video, (2) drawing one bounding box around the hand, and (3) drawing one line for the long axis of the hand. Steps (2) and (3) involve drawing just two items (one box and one line) for a single frame of the video, and even this process could be automated in future work. We would expect our method to detect any oscillatory movement that is visible in the video, regardless of distance of the camera from the body part, because the technique is a measure of pixel movement. Purposefully, we did not tightly constrain the distance between the hand and camera as we recognised that flexibility in this parameter would be important in a clinical setting. If hand pixels are moving in an oscillatory manner, the pixel movement will be detected and the frequency of that oscillation measured.

Contactless tremor measurement using optical flow has recently been described using the (markerless) Microsoft Kinect 3D camera system [190]. However, this method is limited by the need for specialist costly hardware, and inaccuracy in tracking smaller movements, such that tremor smaller than 2cm cannot be reliably detected [190].

Three previous studies reported computer processing of standard video to measure tremor. In 2013, Hemm-Ode et al applied an optical flow algorithm to videos of two patients with hand tremor during Deep Brain Stimulation (DBS) and found “similar trends” in optical flow and accelerometer [191]. However, the authors did not report any quantitative (Hz) measurement of tremor frequency and no summary statistics were provided. A second study measured change in one-dimension, by sampling pixel colour oscillation at static points to give tremor frequency, and showed results comparable with ours (<0.5 Hz difference in 94% of samples) [192]. The advantage of our technique that tracks pixel movement in two dimensions is that it allows potential future measurement of direction, magnitude change, and movement beyond simple tremor. The third previous publication in this area described measurement of tremor during superimposed large amplitude movement [180]. However, it used a more complex technique than ours that required human video labelling of multiple frames to train the computer to reliably detect hands, and agreement with accelerometer was considerably lower than our results (mean absolute error of 1.066 Hz for postural tremor and 1.253 Hz for rest tremor).

The tremor frequency study reported in this chapter has several limitations. Although 37 distinct videos were used (varying by participant, hand, rest/posture), the participant group was small, so we cannot yet be sure that the findings would generalise to a wider population. Although no significant inverse correlation was found between error and amplitude, we cannot rule out the possibility that a significant correlation might emerge in a larger or more diverse sample. Measurement of tremor frequency is most useful when paired with electromyography [85, 88], and camera-based computer vision cannot yet measure muscle contraction. Indeed, determination of tremor frequency alone is of limited clinical value, because the typical frequency ranges of most different tremor types overlap [73, 86]. However, the change in tremor frequency with activation procedures such as weight loading can be more useful to diagnose tremor, which could be added to our protocol without a need for new computing techniques. In addition, there are some situations where measurement of tremor frequency can be useful in itself, e.g. by demonstrating variation of tremor frequency over time in functional tremor [86], a potential future application of our technique (combined with Short-time Fourier Transform) [193]. A camera measurement is two dimensional and, similar to single axis (one dimensional) clinical accelerometer, movements exactly perpendicular to the camera angle potentially may not be recorded, but this could be rectified by moving the camera to a different angle [86]. Two-dimensional video cannot measure absolute tremor amplitude but in principle the relative amplitude (e.g. in relation to finger width), or change in amplitude over time, could be derived from pixel movement in future work. We have not quantified the lower limits of amplitude for detection of tremor frequency using our method, but optical flow is known to be sensitive to small amplitude movement in standard video (provided there is not additional movement 'noise' in the background of the video frame), e.g. [194]. It is notable that 15 of 37 videos were rated as <1cm amplitude, and we did not find any significant relationship between tremor amplitude and video measure accuracy in our current sample. Finally, a tripod is not available in ordinary clinical settings, but smartphone cameras use image stabilisation software and in the future our method could include labelling of a static background reference point, or combination with hand tracking algorithms [195].

For several recordings, the frequency distribution showed two distinct peaks of similar height (power), with each being slightly higher than the other in video vs accelerometer measures. This is consistent with the recognition that tremor can have more than one contributing component with distinct frequencies [86]. However, to allow simple comparison of single number results, we selected the highest frequency peak in these situations. A future approach could perhaps provide a two frequency result to provide more detailed clinical information

Summary

In summary, hands that appear still in the original video show an appearance of parkinsonian tremor after Eulerian magnification for some videos in both Parkinson's patients and controls. At the group level there is more tremor seen in the Parkinson's patients, and clinician classification of hands as Parkinsonian or control is improved after amplification. This group-level effect does not have a direct clinical use, and may largely represent the amplification of 'noise' or physiological tremor in both groups, but it suggests a subclinical Parkinson's tremor component is present within the amplified signal.

For the measurement of tremor frequency, we have described a simple method to measure hand tremor frequency from a 60 second smartphone video that shows good agreement with accelerometer measurements and has the potential to provide a ‘point and press’ contactless measure of tremor frequency within standard clinical settings.

Chapter 6, Discussion

This thesis has demonstrated that computer vision applied to smartphone video can derive meaningful clinical measures of finger tapping bradykinesia and hand tremor.

In **Chapter 2**, I summarised previous publications reporting technological methods to record the movements of the finger tapping test for bradykinesia. Almost all of the approaches are limited by a requirement for hardware or a requirement for patients to interact with a smartphone app. They are also notable for highly variable results across studies, whether comparing group means, classifying results into Parkinson's and control, or testing correlation with clinical ratings. In **Chapter 3**, movement disorder experts showed considerable disagreement when judging one of the most common clinical tests for bradykinesia, the finger tapping test. Intraclass correlation coefficients were moderate, and around a quarter of control hand videos were judged as showing bradykinesia. **Chapter 4** described two different methods to record the movement of the finger tapping test for bradykinesia in video. Using overall hand movement measured by pixel optical flow, the classification of low versus high bradykinesia showed test accuracy of 0.8, and the classification into Parkinson's and control hands showed test accuracy of 0.67. Using the DeepLabDut pose estimation algorithm, there were good correlations between computer vision video measures of bradykinesia and clinical ratings (-0.70, 0.65, -0.61, -0.66, 0.56, -0.50, all $p<0.001$)). **Chapter 5** applied computer vision to hand tremor. A technique to amplify small movement in video revealed apparent subclinical tremor in ostensibly non-tremulous hands from both Parkinson's and control participants, but a significantly higher proportion of Parkinson's hands were correctly classified by clinicians as 'parkinsonian' after movement amplification. **Chapter 5** also described a technique to measure the dominant frequency of hand tremor from video using optical flow, showing agreement to within 0.5 Hz compared with a 'gold standard' accelerometer recording.

The importance of these results is that they provide proof of concept for a clinical tool that could be used to augment or to automate the clinical judgement of the signs of neurological disease. In contrast to other approaches, such a tool would not require special hardware or patient engagement with a new behaviour. Cameras and computers are ubiquitous, available in almost all clinical encounters. The only patient engagement required is that which already occurs, namely participation in the neurological examination, for example tapping finger and thumb together, or permitting observation of tremulous hands in several positions.

There are specific limitations to the work in each chapter and these have been described in those chapters. However, there are also some overall limitations that run through the thesis.

A single monocular camera cannot measure distance or depth. Since human movement takes place in three dimensions, that leaves some movement unrecorded. Furthermore, measurements of movement that rely on distance, such as amplitude and speed, can only be relative rather than absolute. This is likely to be a fairly minor issue in relation to finger tapping, because the movement occurs largely in two dimensions and distance can be normalised to maximum finger thumb amplitude. The results of **Chapter 4** suggest that a monocular camera captures most of the useful finger tapping information in two dimensions, with results comparable with devices that measure three-dimensional movement. However, for tremor, the standard clinical rating scales, used as research outcome measures, are clinician grading of tremor amplitude (in centimetres), not measures of tremor frequency. That is true for both Parkinson's [8] and Essential Tremor [196]. This clinical information cannot be captured by monocular camera computer vision. While the results of **Chapter 5** show that monocular video can record tremor frequency, the most common tremor conditions overlap in their frequency ranges, so frequency in isolation has limited clinical uses (e.g., a change in frequency over time can help to diagnose functional tremor [197]). There are potential solutions for this without special equipment. One would be to derive a relative measure of distance for tremor, similar to using maximum finger-thumb opening for finger tapping. For example, distance could be normalised to index finger width. Another is to use two monocular cameras, e.g., two smartphones, because pose estimation algorithms can then construct 3D coordinates, with one network trained on multiple views, provided the relative camera positions are calibrated [116]. That would not be easy to implement: coordinated, calibrated, and linked use of two smartphones is not an everyday practice. An alternative possibility would be to use data collected by the smartphone depth sensing camera. Most current smartphone models now have infrared array 'LiDAR' (light detection and ranging) infrared scanners linked to their front facing camera that can measure depth

[198]. These are used as the biometric facial recognition system to unlock the iPhone [199, 200]. In theory, these depth sensors that are already integrated with standard smartphone cameras could be used to measure tremor amplitude during video in a contactless manner, without special equipment.

In clinical practice, individual neurological examination signs are not considered in isolation. They form part of an overall judgement, both in diagnosis and also in assessment of progression or response to treatment. This overall judgement includes multiple examination signs, but also the history of the patient's symptoms. There is a medical aphorism that a careful history of symptoms "will lead to the diagnosis 80% of the time" [201]. This is based upon a 1975 study that asked clinicians to give differential diagnoses at three different points in the assessment process [202], a finding broadly replicated in subsequent studies [201, 203]. As such, an automated, objective measure of bradykinesia or tremor is limited in the clinical information it can provide, and limited in the degree to which it can augment clinician judgement. A large number of examination signs are visual, so that it would be possible to apply computer vision to a much broader neurological examination, then combine this with automated measurement of symptoms (in other words collect data from specific questions about symptoms). It would then be theoretically possible to aim to automate diagnosis or overall assessment, potentially with an accuracy that approaches that of expert clinicians, or perhaps exceeds their accuracy, when compared against some ultimate gold standard, such as post-mortem studies. However, when machine learning techniques are applied to multiple features in a classification task, there is a high risk of 'overfitting' to the data. In other words, classification is based upon variation in the specific dataset, but does not generalise well to data in general. To minimise that would require a very large set of data (patient videos), difficult to obtain in practice. An alternative is to give *a priori* weighting or rules for different symptoms and signs according to clinical knowledge, for example, the symptom of rest tremor is more common in Parkinson's, while kinetic tremor is more common in Essential Tremor. This would be the application of a Bayesian statistical approach [141, 204]. However, the exact relative weighting to give such clinical details would inevitably be fairly arbitrary, running the risk of replicating existing clinician bias or error.

Another limitation of computer vision for neurological examination is the lack of a reliable 'ground truth' or benchmark for many clinical signs. **Chapter 3** demonstrated that expert clinicians show only moderate intraclass correlation for the judgement of bradykinesia severity. To some extent this can be mitigated by using a large number of raters, as with bradykinesia correlations in **Chapter 4**. However, as described in **Chapter 1** the modern concept of bradykinesia only dates from the 1980s [72]. It is a much more recent idea than the more general one of Parkinson's itself [67]. Clinician disagreement might occur because bradykinesia partly exists in the mind of the clinician rather than the movement of the patient. If a clinical sign is to some extent an artificial construct, then to that extent it cannot be objectively measured. Correlation with clinical ratings is not necessarily a good test of technology measures.

In addition to the lack of reliable ground truth for the presence or severity of clinical signs, there is a similar problem in relation to overall diagnosis based on a broader assessment. Tremor frequency can be objectively measured beyond clinician judgement, and that is also theoretically possible for tremor amplitude. However, interpretation of those patterns of frequency and relative amplitude to make a diagnosis involves an element of clinician judgement. This is because there are no entirely sensitive and specific patterns for different diagnoses or prognoses [86, 98, 205, 206]. For example, entrainment of tremor frequency to contralateral limb tapping frequency is typical of the diagnosis functional tremor [197] but it is not always seen in that condition, and can occur in other diagnoses [206].

A possible solution to unreliable clinician ground truths is to assess computer vision measures against gold standards that are not based on clinician judgement. This is really only possible for diagnosis rather than severity. For example, a dopamine transporter ('DaT') scan measures loss of dopaminergic neurons in the striatum, and can distinguish neurodegenerative parkinsonian syndromes from other causes of parkinsonism or tremor, such as drug-induced parkinsonism, essential tremor, dystonic tremor. The ultimate gold standard for Parkinson's diagnosis is post-mortem. Any study involving such gold standards can only involve low numbers of patients (DaT scans involve ionising radiation and cost, while only a small minority of patients undergo post-mortem). This is not compatible with the need for large datasets in machine learning studies that involve multiple features (multiple examination signs and patient reported symptoms).

One potential diagnostic tool that is universally present is the passage of time. The accuracy of diagnosis is lower early in the course of a disease, but improves later, as existing clinical features either progress and become more pronounced, or do not, and as new clinical features develop, or do not. Similarly, judgement of change over time is easier as more time passes and repeated assessments take place, making trends clearer. One possible future approach to testing computer vision measures is to compare them against future clinical assessments (diagnosis), for example two years after the computer vision measures were recorded. An example is distinguishing the signs of essential tremor from early Parkinson’s disease. At follow-up assessments, the diagnosis often becomes clearer for a clinician, and could be used as a benchmark against which to test computer vision measures taken at the original presentation.

When testing technology to recognise a ‘diagnosis’, for example Parkinson’s disease, it is common to test the categorisation of disease versus control (no disease). However, in reality, neurologists must distinguish between different disease diagnoses rather than disease versus no disease. A patient is referred to the outpatient clinic with symptoms, and the neurologist tasked with diagnosing them. Technology to augment or automate clinician diagnosis should be tested for the ability to distinguish common differential diagnoses for a given symptom, e.g. causes of tremor or causes of gait impairment.

Another possible approach for computer vision to assess neurological signs would be to apply a deep learning convolutional neural network directly to video, to learn movement patterns that discriminate Parkinson’s from control, or different grades of severity, without predefined rules. This would be a three-dimensional convolutional neural network (two dimensions of pixels and one dimension of time). It is possible that there are movement patterns present during finger tapping in Parkinson’s that are outside the current definition of bradykinesia. Such patterns could be invisible to clinicians, or only recognisable to clinicians in an automatic, unconscious way. Deep learning using neural networks is a machine learning technique that does not require human-defined features, such as finger to thumb distance or opening velocity; instead, it can learn the features most predictive of a given category, without predefined rules. This provides a way to look for patterns that are characteristic of Parkinson's in video, or characteristic of different grades of bradykinesia, without restricting patterns to one-dimensional, human-defined measures, such as opening velocity. That may find new, additional, or better features to discriminate and measure Parkinson's.

The requirement for clinician labelling to provide the ground truth for the severity of clinical signs (e.g. bradykinesia grading) makes it difficult to obtain a large dataset to allow machine learning techniques. It would not be easy to obtain thousands of clinician ratings of videos. In theory, a possible solution for this is the technique of 'self-supervised learning', in which a deep neural network is asked "to predict one part of the input data – or a label pragmatically derivable thereof – given another part of the input" [207]. The technique removes the requirement for large annotated datasets.

Conclusion

There is a history over many years of publications reporting technology to measure bradykinesia and tremor. This technology has been limited by a requirement for either special equipment or patient engagement with a smartphone app, and results of the application of such technology to bradykinesia are variable and inconsistent. This thesis has demonstrated that expert rating of finger tapping bradykinesia shows only moderate interrater agreement, suggesting the clinical sign exists to some extent in the eye of the clinician rather than purely the hand of the patient. The results presented in this thesis demonstrate that standard monocular smartphone video can be used to derive measures of finger tapping bradykinesia and tremor frequency that correlate well with clinical ratings and accelerometer measurement respectively. The findings suggest that the smartphone camera could be used to augment clinical assessment.

References

1. Innes, J.A., A.R. Dover, and K. Fairhurst, *Macloed's Clinical Examination, 14th Edition*. 2019: Elsevier.
2. Patten, J.P., *Neurological Differential Diagnosis*. 1996: Springer Berlin Heidelberg.
3. Brazis, P.W., J.C. Masdeu, and J. Biller, *Localization in Clinical Neurology*. 8th ed. 2021: Lippincott Williams and Wilkins.
4. Campbell, B.C.V. and P. Khatri, *Stroke.* The Lancet, 2020. **396**(10244): p. 129-142.
5. Thijs, R.D., et al., *Epilepsy in adults.* The Lancet, 2019. **393**(10172): p. 689-701.
6. Brott, T., et al., *Measurements of acute cerebral infarction: a clinical examination scale.* Stroke, 1989. **20**(7): p. 864-870.
7. Postuma, R.B., et al., *MDS clinical diagnostic criteria for Parkinson's disease.* Movement Disorders, 2015. **30**(12): p. 1591-1601.
8. Goetz, C.G., et al., *Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results.* Movement Disorders, 2008. **23**(15): p. 2129-2170.
9. Thompson, A.J., et al., *Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria.* The Lancet Neurology, 2018. **17**(2): p. 162-173.
10. Kurtzke, J.F., *Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS).* Neurology, 1983. **33**(11): p. 1444-1452.
11. Parkinson, J., *An Essay on the Shaking Palsy.* 1817.
12. Giovannoni, G., et al., *Bradykinesia akinesia inco-ordination test (BRAIN TEST): an objective computerised assessment of upper limb motor function.* Journal of neurology, neurosurgery, and psychiatry, 1999. **67**(5): p. 624-629.
13. Homann, C.N., et al., *The Bradykinesia Akinesia Incoordination Test (BRAIN TEST), an objective and user-friendly means to evaluate patients with parkinsonism.* Movement Disorders, 2000. **15**(4): p. 641-647.
14. Pal, P.K., et al., *Alternating two finger tapping with contralateral activation is an objective measure of clinical severity in Parkinson's disease and correlates with PET [18F]-DOPA Ki.* Parkinsonism and Related Disorders, 2001. **7**(4): p. 305-309.
15. Tavares, A.L.T., et al., *Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation.* Movement Disorders, 2005. **20**(10): p. 1286-1298.
16. Papapetropoulos, S., et al., *Objective monitoring of tremor and bradykinesia during DBS surgery for Parkinson disease.* Neurology, 2008. **70**(15): p. 1244-1249.
17. Papapetropoulos, S., et al., *Objective quantification of neuromotor symptoms in parkinson's disease: Implementation of a portable, computerized measurement tool.* Parkinson's Disease, 2010: p. 760196-760196.
18. Maetzler, W., et al., *Digitomotography in Parkinson's disease: a cross-sectional and longitudinal study.* PloS one, 2015. **10**(4): p. e0123914-e0123914.
19. Kassavetis, P., et al., *Developing a Tool for Remote Digital Assessment of Parkinson's Disease.* MOVEMENT DISORDERS CLINICAL PRACTICE, 2016. **3**(1): p. 59-64.
20. Mitsi, G., et al., *Biometric digital health technology for measuring motor function in Parkinson's disease: Results from a feasibility and patient satisfaction study.* Frontiers in Neurology, 2017. **8**(JUN): p. 273-273.

21. Lalvay, L., et al., *Quantitative Measurement of Akinesia in Parkinson's Disease.* Movement Disorders Clinical Practice, 2017. **4**(3): p. 316-322.
22. Prince, J., S. Arora, and M. de Vos, *Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes.* PHYSIOLOGICAL MEASUREMENT, 2018. **39**(4).
23. Roalf, D.R., et al., *Quantitative assessment of finger tapping characteristics in mild cognitive impairment, Alzheimer's disease, and Parkinson's disease.* Journal of Neurology, 2018. **265**(6): p. 1365-1375.
24. Yokoe, M., et al., *Opening velocity, a novel parameter, for finger tapping test in patients with Parkinson's disease.* Parkinsonism and Related Disorders, 2009. **15**(6): p. 440-444.
25. Costa, J., et al., *Nonlinear dynamic analysis of oscillatory repetitive movements in Parkinson's disease and essential tremor.* Movement Disorders, 2010. **25**(15): p. 2577-2586.
26. Kim, J.-W., et al., *Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with Parkinson's disease.* Medical & biological engineering & computing, 2011. **49**(3): p. 365-371.
27. Heldman, D.A., et al., *The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures.* Movement Disorders, 2011. **26**(10): p. 1859-63.
28. Stamatakis, J., et al., *Finger tapping clinimetric score prediction in Parkinson's disease using low-cost accelerometers.* Computational intelligence and neuroscience, 2013. **2013**: p. 717853-717853.
29. Heldman, D.A., et al., *Clinician versus machine: Reliability and responsiveness of motor endpoints in Parkinson's disease.* Parkinsonism and Related Disorders, 2014. **20**(6): p. 590-595.
30. Lee, M.J., et al., *Kinematic Analysis in Patients with Parkinson's Disease and SWEDD.* JOURNAL OF PARKINSONS DISEASE, 2014. **4**(3): p. 421-430.
31. Kim, J.-W., et al., *Regression models for the quantification of Parkinsonian bradykinesia.* Bio-medical materials and engineering, 2015. **26**(Supplement 1): p. S2249-S2258.
32. Martinez-Manzanera, O., et al., *A Method for Automatic and Objective Scoring of Bradykinesia Using Orientation Sensors and Classification Algorithms.* IEEE Transactions on Biomedical Engineering, 2016. **63**(5): p. 1016-1024.
33. Heldman, D.A., et al., *App-Based Bradykinesia Tasks for Clinic and Home Assessment in Parkinson's Disease: Reliability and Responsiveness.* Journal of Parkinson's disease, 2017. **7**(4): p. 741-747.
34. di Biase, L., et al., *Quantitative analysis of bradykinesia and rigidity in Parkinson's disease.* Frontiers in Neurology, 2018. **9**(MAR): p. 1-12.
35. Agostino, R., et al., *Impairment of individual finger movements in Parkinson's disease.* Movement Disorders, 2003. **18**(5): p. 560-565.
36. Lainscsek, C., et al., *Finger tapping movements of Parkinson's disease patients automatically rated using nonlinear delay differential equations.* Chaos, 2012. **22**: p. 013119-013119.
37. Krupicka, R., et al., *Motion capture system for finger movement measurement in Parkinson disease.* Radioengineering, 2014. **23**(2): p. 659-664.

38. Ruzicka, E., et al., *Tests of manual dexterity and speed in Parkinson's disease: Not all measure the same.* Parkinsonism & related disorders, 2016. **28**: p. 118-123.
39. Bologna, M., et al., *Bradykinesia in early and advanced Parkinson's disease.* Journal of the Neurological Sciences, 2016. **369**: p. 286-291.
40. Bank, P.J.M., et al., *Optical Hand Tracking: A Novel Technique for the Assessment of Bradykinesia in Parkinson's Disease.* Movement Disorders Clinical Practice, 2017. **4**(6): p. 875-883.
41. Kandori, A., et al., *Quantitative magnetic detection of finger movements in patients with Parkinson's disease.* Neuroscience Research, 2004. **49**(2): p. 253-260.
42. Shima, K., et al., *Measurement and Evaluation of Finger Tapping Movements Using Log-linearized Gaussian Mixture Networks.* SENSORS, 2009. **9**(3): p. 2187-2201.
43. Sano, Y., et al., *Quantifying Parkinson's disease finger-tapping severity by extracting and synthesizing finger motion properties.* Medical and Biological Engineering and Computing, 2016. **54**(6): p. 953-965.
44. Gao, C., et al., *Objective assessment of bradykinesia in Parkinson's disease using evolutionary algorithms: Clinical validation.* Translational Neurodegeneration, 2018. **7**(1): p. 1-8.
45. Teo, W.P., et al., *Comparing kinematic changes between a finger-tapping task and unconstrained finger flexion-extension task in patients with Parkinson's disease.* Experimental brain research, 2013. **227**(3): p. 323-331.
46. Lee, C.Y., et al., *A Validation Study of a Smartphone-Based Finger Tapping Application for Quantitative Assessment of Bradykinesia in Parkinson's Disease.* PLOS ONE, 2016. **11**(7).
47. Ling, H., et al., *Hypokinesia without decrement distinguishes progressive supranuclear palsy from Parkinson's disease.* Brain, 2012. **135**(4): p. 1141-1153.
48. Adler, C.H., et al., *Low clinical diagnostic accuracy of early vs advanced Parkinson disease.* Neurology, 2014. **83**(5): p. 406 LP-412.
49. Goedert, M., et al., *100 years of Lewy pathology.* Nature Reviews Neurology, 2013. **9**(1): p. 13-24.
50. Shahmoradian, S.H., et al., *Lewy pathology in Parkinson’s disease consists of crowded organelles and lipid membranes.* Nature Neuroscience, 2019. **22**(7): p. 1099-1109.
51. Obeso, J.A., et al., *The basal ganglia in Parkinson's disease: Current concepts and unexplained observations.* Annals of Neurology, 2008. **64**(SUPPL. 2): p. 30-46.
52. Blesa, J., et al., *Motor and non-motor circuit disturbances in early Parkinson disease: which happens first?* Nature Reviews Neuroscience, 2022. **23**(2): p. 115-128.
53. Tinkhauser, G., et al., *Beta burst dynamics in Parkinson's disease off and on dopaminergic medication.* Brain, 2017. **140**(11): p. 2968-2981.
54. Braak, H., et al., *Staging of brain pathology related to sporadic Parkinson's disease.* Neurobiology of Aging, 2003. **24**(2): p. 197-211.
55. Henderson, L., et al., *Scales for rating motor impairment in Parkinson's disease: Studies of reliability and convergent validity.* Journal of Neurology Neurosurgery and Psychiatry, 1991. **54**(1): p. 18-24.
56. Kishore, A., et al., *Unilateral versus bilateral tasks in early asymmetric Parkinson's disease: Differential effects on bradykinesia.* Movement Disorders, 2007. **22**(3): p. 328-333.
57. Bloem, B.R., M.S. Okun, and C. Klein, *Parkinson's disease.* The Lancet, 2021. **397**(10291): p. 2284-2303.

58. Marras, C. and A. Lang, *Parkinson's disease subtypes: Lost in translation?* Journal of Neurology, Neurosurgery and Psychiatry, 2013. **84**(4): p. 409-415.

59. Delenclos, M., et al., *Biomarkers in Parkinson's disease: Advances and strategies.* Parkinsonism and Related Disorders, 2016. **22**: p. S106-S110.

60. Heinzel, S., et al., *Update of the MDS research criteria for prodromal Parkinson's disease.* Movement Disorders, 2019: p. mds.27802-mds.27802.

61. Hawkes, C.H., *The prodromal phase of sporadic Parkinson's disease: Does it exist and if so how long is it?* Movement Disorders, 2008. **23**(13): p. 1799-1807.

62. Feigin, V.L., et al., *Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016.* The Lancet Neurology, 2019. **18**(5): p. 459-480.

63. Noyce, A.J., et al., *Meta-analysis of early nonmotor features and risk factors for Parkinson disease.* Annals of Neurology, 2012. **72**(6): p. 893-901.

64. Singleton, A.B. and H. Houlden, *The Genetics and neuropathology of PD.* Acta neuropathologica, 2013. **124**(3): p. 325-338.

65. Blauwendraat, C., M.A. Nalls, and A.B. Singleton, *The genetic architecture of Parkinson's disease.* The Lancet Neurology, 2020. **19**(2): p. 170-178.

66. Zhang, Z.X., Z.H. Dong, and G.C. Román, *Early descriptions of Parkinson disease in ancient China.* Archives of Neurology, 2006. **63**(5): p. 782-784.

67. Goetz, C.G., *The history of Parkinson's disease: Early clinical descriptions and neurological therapies.* Cold Spring Harbor Perspectives in Medicine, 2011. **1**(1).

68. Charcot, J.M., *Lectures on Diseases of the Nervous system: Delivered at La Salpêtrière: Translated by G. Sigerson*. 1877, London: The New Sydenham Society. 144-145.

69. Schilder, J.C.M., et al., *The terminology of akinesia, bradykinesia and hypokinesia: Past, present and future.* Parkinsonism and Related Disorders, 2017. **37**: p. 27-35.

70. Hornykiewicz, O., *Dopamine miracle: From brain homogenate to dopamine replacement.* Movement Disorders, 2002. **17**(3): p. 501-508.

71. Gibb, W.R.G. and A.J. Lees, *Occasional review: the relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease.* J Neurol Neurosurg Psychiatry, 1988. **51**: p. 745-752.

72. Fahn, S. and R.L. Elton, *Unified Parkinson's Disease Rating Scale*, S. Fahn, et al., Editors. 1987, NJ. Macmillan Health Care Information: Florham Park. p. 153-164.

73. Bhatia, K.P., et al., *Consensus Statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society.* Movement Disorders, 2018. **33**(1): p. 75-87.

74. Alty, J.E. and P.A. Kempster, *A practical guide to the differential diagnosis of tremor.* Postgraduate Medical Journal, 2011. **87**(1031): p. 623-629.

75. Zach, H., et al., *The clinical evaluation of Parkinson's tremor.* Journal of Parkinson's Disease, 2015. **5**(3): p. 471-474.

76. Helmich, R.C., *The cerebral basis of Parkinsonian tremor: A network perspective.* Movement Disorders, 2018. **33**(2): p. 219-231.

77. Wilken, M., et al., *Sensitivity and specificity of different hand positions to assess upper limb rest tremor.* Movement Disorders, 2019. **34**(4): p. 575-579.

78. Zaidel, A., et al., *Akineto-rigid vs. tremor syndromes in Parkinsonism.* Current Opinion in Neurology, 2009. **22**(4): p. 387-393.

79. Louis, E.D. and R. Ottman, *How many people in the USA have essential tremor? Deriving a population estimate based on epidemiological data.* Tremor and other hyperkinetic movements (New York, N.Y.), 2014. **4**: p. 259-259.

80. Barrantes, S., et al., *Differential diagnosis between Parkinson's disease and essential tremor using the smartphone's accelerometer.* PLoS ONE, 2017. **12**(8): p. 1-12.

81. Algarni, M. and A. Fasano, *The overlap between Essential tremor and Parkinson disease.* Parkinsonism and Related Disorders, 2018. **46**: p. S101-S104.

82. Nahab, F.B., E. Peckham, and M. Hallett, *Essential tremor, deceptively simple.* Practical Neurology, 2007. **7**(4): p. 222-233.

83. Louis, E.D., et al., *Essential tremor-plus: a controversial new concept.* The Lancet Neurology, 2020. **19**(3): p. 266-270.

84. Rajalingam, R., et al., *Essential tremor plus is more common than essential tremor: Insights from the reclassification of a cohort of patients with lower limb tremor.* Parkinsonism and Related Disorders, 2018. **56**(May): p. 109-110.

85. Hess, C.W. and S.L. Pullman, *Tremor: clinical phenomenology and assessment techniques.* Tremor and other hyperkinetic movements (New York, N.Y.), 2012. **2**: p. 1-15.

86. Vial, F., et al., *How to do an electrophysiological study of tremor.* Clinical Neurophysiology Practice, 2019. **4**: p. 134-142.

87. Hopfner, F. and R.C. Helmich, *The etiology of essential tremor: Genes versus environment.* Parkinsonism and Related Disorders, 2018. **46**: p. S92-S96.

88. Elble, R.J. and J. McNames, *Using portable transducers to measure tremor severity.* Tremor and Other Hyperkinetic Movements, 2016. **2016**: p. 1-12.

89. Joundi, R.A., et al., *Rapid tremor frequency assessment with the iPhone accelerometer.* Parkinsonism and Related Disorders, 2011. **17**(4): p. 288-290.

90. Louis, E.D., *The evolving definition of essential tremor: What are we dealing with?* Parkinsonism and Related Disorders, 2018. **46**: p. S87-S91.

91. Jain, S., S.E. Lo, and E.D. Louis, *Common Misdiagnosis of a Common Neurological Disorder.* Archives of Neurology, 2006. **63**(8): p. 1100-1100.

92. Bajaj, N.P.S., et al., *Accuracy of clinical diagnosis in tremulous parkinsonian patients: a blinded video study.* Journal of Neurology, Neurosurgery and Psychiatry, 2010. **81**: p. 1223-1228.

93. Adler, C.H., et al., *Low clinical diagnostic accuracy of early vs advanced Parkinson disease.* Neurology, 2014. **83**(83): p. 406-412.

94. Rizzo, G., et al., *Accuracy of clinical diagnosis of Parkinson's disease: A Systematic Review and Meta-Analysis.* Neurology, 2016. **87**(2): p. 237-8.

95. Goetz, C.G. and G.T. Stebbins, *Assuring interrater reliability for the UPDRS motor section: Utility of the UPDRS teaching tape.* Movement Disorders, 2004. **19**(12): p. 1453-1456.

96. Bennett, D.A., et al., *Metric properties of nurses' ratings of parkinsonian signs with a modified Unified Parkinson's Disease Rating Scale.* Neurology, 1997. **49**(6): p. 1580-1587.

97. Kostikis, N., et al., *A smartphone-based tool for assessing parkinsonian hand tremor.* IEEE Journal of Biomedical and Health Informatics, 2015. **19**(6).

98. Elble, R.J. and W. Ondo, *Tremor rating scales and laboratory tools for assessing tremor.* Journal of the Neurological Sciences, 2022. **435**(August 2021): p. 120202-120202.

99. Zhan, A., et al., *Using smartphones and machine learning to quantify Parkinson disease severity the mobile Parkinson disease score.* JAMA Neurology, 2018. **75**(7): p. 876-880.

100. Arora, S., et al., *Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study.* Parkinsonism and Related Disorders, 2015. **21**(6): p. 650-653.

101. Shapiro, L.G. and G. Stockman, *Computer vision*. 2001: Prentice Hall.

102. Szeliski, R., *Computer vision: Algorithms and applications*. 2011. 812-812.

103. Bojarski, M., et al., *Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car.* arXiv preprint arXiv:1704.07911, 2017: p. 1-8.

104. Huval, B., et al., *An Empirical Evaluation of Deep Learning on Highway Driving.* arXiv preprint arXiv:1504.01716, 2015: p. 1-7.

105. Buch, N., S.A. Velastin, and J. Orwell, *A review of computer vision techniques for the analysis of urban traffic.* IEEE Transactions on Intelligent Transportation Systems, 2011. **12**(3): p. 920-939.

106. Jacques, J.C.S., S.R. Mussef, and C.R. Jung, *Crowd analysis using computer vision techniques.* IEEE Signal Processing Magazine, 2010. **27**(5): p. 66-77.

107. Masi, I., et al., *Deep Face Recognition: A Survey.* Proceedings - 31st Conference on Graphics, Patterns and Images, SIBGRAPI 2018, 2019: p. 471-478.

108. Horn, B.K. and B.G. Schunck, *Determining Optical Flow.* Artificial Intelligence, 1981. **17(1-3)**: p. 183-203.

109. Chaudhry, R., et al., *Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions.* 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, 2009. **2009 IEEE**: p. 1932-1939.

110. Happy, S.L. and A. Routray, *Fuzzy Histogram of Optical Flow Orientations for Micro-expression Recognition.* IEEE Transactions on Affective Computing, 2017. **3045**(c): p. 1-13.

111. Gong, W., et al., *Human pose estimation from monocular images: A comprehensive survey.* Sensors (Switzerland), 2016. **16**(12): p. 1-39.

112. Insafutdinov, E., et al., *Deepercut: A deeper, stronger, and faster multi-person pose estimation model.* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016. **9910 LNCS**: p. 34-50.

113. Gulshan, V., et al., *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.* JAMA - Journal of the American Medical Association, 2016. **316**(22): p. 2402-2410.

114. Lecun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.

115. Deng, J., et al. *ImageNet: A Large-Scale Hierarchical Image Database*. IEEE.

116. Nath, T., et al., *Using DeepLabCut for 3D markerless pose estimation across species and behaviors.* Nature Protocols, 2019. **14**(7): p. 2152-2176.

117. Mathis, A., et al., *DeepLabCut: markerless pose estimation of user-defined body parts with deep learning.* Nature Neuroscience, 2018. **21**(9): p. 1281-1289.

118. Wu, H.Y., et al., *Eulerian video magnification for revealing subtle changes in the world.* ACM Transactions on Graphics, 2012. **31**(4).

119. Jollife, I.T. and J. Cadima, *Principal component analysis: A review and recent developments.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016. **374**(2065).

120. Smith, L.I., *A tutorial on principal components analysis*. 2002.

121. Shlens, J., *A Tutorial on Principal Component Analysis.* 2005.

122. Theobald, O., *Machine learning for absolute beginners: a plain English introduction*. Vol. 157. 2017: Scatterplot press.

123. Leung, K.M., *Naive bayesian classifier.* Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007: p. 123-156.

124. Higham, N.J., et al., *Princeton companion to applied mathematics*. 2015: Princeton University Press.

125. Fourier, J.B.J., *The analytical theory of heat*. 2009: Cambridge University Press. 1-466.

126. Spyers-Ashby, J.M., P.G. Bain, and S.J. Roberts, *A comparison of fast fourier transform (FFT) and autoregressive (AR) spectral estimation techniques for the analysis of tremor data.* Journal of Neuroscience Methods, 1998. **83**(1): p. 35-43.

127. DiVincenzo, N., *What is the Fourier Transform ?* 2019.

128. Mantha, S., et al., *Comparing methods of clinical measurement: Reporting standards for bland and altman analysis.* Anesthesia and Analgesia, 2000. **90**(3): p. 593-602.

129. Bland, M.J. and D.G. Altman, *Measuring agreement in method comparison studies.* Statistical Methods in Medical Research, 1999. **8**: p. 135-60.

130. Martin Bland, J. and D.G. Altman, *Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement.* The Lancet, 1986. **327**(8476): p. 307-310.

131. Camicioli, R., et al., *Discriminating mild parkinsonism: Methods for epidemiological research.* Movement Disorders, 2001. **16**(1): p. 33-40.

132. Hughes, A.J., et al., *Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases.* Journal of Neurology Neurosurgery and Psychiatry, 1992. **55**(3): p. 181-184.

133. Goetz, C.G., et al., *Testing Objective Measures of Motor Impairment in Early Parkinson's Disease: Feasibility Study of an At-Home Testing Device.* MOVEMENT DISORDERS, 2009. **24**(4): p. 551-556.

134. van den Noort, J.C., et al., *Quantification of Hand Motor Symptoms in Parkinson’s Disease: A Proof-of-Principle Study Using Inertial and Force Sensors.* Annals of Biomedical Engineering, 2017. **45**(10): p. 2423-2436.

135. Liu, Y., et al., *Vision-Based Method for Automatic Quantification of Parkinsonian Bradykinesia.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019. **27**(10): p. 1952-1961.

136. Lee, W., et al., *Objective evaluation of bradykinesia in Parkinson's disease using an inexpensive marker-less motion tracking system.* Physiological measurement, 2019. **40**(1): p. 14004-14004.

137. Edwards, S., et al., *Caution using data from triaxial accelerometers housed in player tracking units during running.* Journal of Sports Sciences, 2019. **37**(7): p. 810-818.

138. Webster, D. and O. Celik, *Systematic review of Kinect applications in elderly care and stroke rehabilitation.* Journal of NeuroEngineering and Rehabilitation, 2014. **11**(108): p. 1-24.

139. Galna, B., et al., *Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease.* GAIT & POSTURE, 2014. **39**(4): p. 1062-1068.
140. Kahneman, D., *Thinking, fast and slow*. 2011: Macmillan.
141. Vancheri, F., *Bayesian principles or Gestalt perception for clinical judgment.* Internal and Emergency Medicine, 2015. **10**(2): p. 253-253.
142. Cervellin, G., L. Borghi, and G. Lippi, *Do clinicians decide relying primarily on Bayesians principles or on Gestalt perception? Some pearls and pitfalls of Gestalt perception in medicine.* Internal and Emergency Medicine, 2014. **9**(5): p. 513-519.
143. Hamlyn, D.W., *The psychology of perception: A philosophical examination of Gestalt theory and derivative theories of perception (Vol. 13)*. 2017: Routledge.
144. Flood, M.W., et al., *Quantitative clinical assessment of motor function during and following LSVT-BIG® therapy.* Journal of NeuroEngineering and Rehabilitation, 2020. **17**(1): p. 1-19.
145. Williams, S., et al., *Supervised classification of bradykinesia in Parkinson's disease from smartphone videos.* Artificial Intelligence in Medicine, 2020. **110**: p. 101966-101966.
146. Williams, S., et al., *The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia?* Journal of the Neurological Sciences, 2020. **416**: p. 117003-117003.
147. Taylor Tavares, A.L., et al., *Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation.* Movement disorders : official journal of the Movement Disorder Society, 2005. **20**(10): p. 1286-1298.
148. Georgiou, N., et al., *Reduction in external cues and movement sequencing in Parkinson's disease.* Journal of Neurology, Neurosurgery, and Psychiatry, 1994. **57**: p. 368-370.
149. Khan, T., et al., *A computer vision framework for finger-tapping evaluation in Parkinson's disease.* Artificial Intelligence in Medicine, 2014. **60**(1): p. 27-40.
150. Shin, J.H., et al., *Objective measurement of limb bradykinesia using a marker-less tracking algorithm with 2D-video in PD patients.* Parkinsonism and Related Disorders, 2020. **81**: p. 129-135.
151. Lee, M.J., et al., *Impact of regional striatal dopaminergic function on kinematic parameters of Parkinson’s disease.* Journal of Neural Transmission, 2015. **122**(5): p. 669-677.
152. Maetzler, W., et al., *Digitomotography in Parkinson's disease: A cross-sectional and longitudinal study.* PLoS ONE, 2015.
153. Liu, Y., et al., *Vision-Based Method for Automatic Quantification of Parkinsonian Bradykinesia.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019. **27**(10): p. 1952-1961.
154. Burkov, A., *The hundred-page machine learning book*. 2019, Quebec City, Canada.
155. Martinez-Martin, P., et al., *Unified Parkinson's disease rating scale characteristics and structure.* Movement Disorders, 1994. **9**(1): p. 76-83.
156. Post, B., et al., *Unified Parkinson's Disease Rating Scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?* Movement Disorders, 2005. **20**(12): p. 1577-1584.

157. Rabey, J.M., et al., *Evaluation of the Short Parkinson's Evaluation Scale: A New Friendly Scale for the Evaluation of Parkinson's Disease in Clinical Drug Trials.* Clinical Neuropharmacology, 1997. **20**(4): p. 322-337.

158. Luiz, L.M.D., et al., *Intra and inter-rater remote assessment of bradykinesia in Parkinson's disease.* Neurología, 2021(xxxx).

159. Palmer, J.L., et al., *Unified Parkinson's Disease Rating Scale-Motor Exam: Inter-rater reliability of advanced practice nurse and neurologist assessments.* Journal of Advanced Nursing, 2010. **66**(6): p. 1382-1387.

160. Landis, J.R. and G.G. Koch, *The Measurement of Observer Agreement for Categorical Data.* Biometrics, 1977. **33**(1): p. 159-159.

161. Van Rossum, G. and F.L. Drake, *Python 3 Reference Manual*. 2009, Scotts Valley, CA: CreateSpace.

162. Koo, T.K. and M.Y. Li, *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.* Journal of Chiropractic Medicine, 2016. **15**(2): p. 155-163.

163. Liddell, T.M. and J.K. Kruschke, *Analyzing ordinal data with metric models: What could possibly go wrong?* Journal of Experimental Social Psychology, 2018. **79**: p. 328-348.

164. Goetz, C.G., et al., *Movement Disorder Society Task Force Report on the Hoehn and Yahr Staging Scale: Status and Recommendations The Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease.* 2004.

165. Cicchetti, D.V., *Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology*. 1994. p. 284-290.

166. Hughes, A.J., et al., *The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service.* BRAIN, 2002. **125**(4): p. 861-870.

167. Bambach, S., et al., *Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions.* Proceedings of the IEEE International Conference on Computer Vision, 2015. **2015 Inter**: p. 1949-1957.

168. Sandler, M., et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*.

169. Huang, J., et al., *TensorFlow object detection API*.

170. Rother, C., V. Kolmogorov, and A. Blake, *"GrabCut" - Interactive Foreground Extraction using Iterated Graph Cuts.* ACM Transactions on Graphics, 2004. **23**(3): p. 309-314.

171. The MathWorks, I., *MATLAB*. 2017: Massachusetts.

172. Electrophysiology, T.F., *Heart rate variability: standards of measurement, physiological interpretation, and clinical use.* Circulation, 1996. **93(5)**: p. 1043-1065.

173. Orphanidou, C. and D. Wong, *Machine Learning Models for Multidimensional Clinical Data*. 2017, Springer: Cham. p. 177-216.

174. Golland, P. and B. Fischl. *Permutation tests for classification: Towards statistical significance in image-based studies*. Springer Berlin Heidelberg.

175. Abadi, M., et al. *TensorFlow: A System for Large-Scale Machine Learning*.

176. Niaf, E., et al., *Kernel-based learning from both qualitative and quantitative labels: Application to prostate cancer diagnosis based on multiparametric MR imaging.* IEEE Transactions on Image Processing, 2014. **23**(3): p. 979-991.

177. *Top Countries / Markets by Smartphone Penetration & Users*.

178. Espay, A.J., et al., *Impairments of speed and amplitude of movement in Parkinson's disease: A pilot study.* Movement Disorders, 2009. **24**(7): p. 1001-1008.

179. Khan, T., J. Westin, and M. Dougherty, *Motion Cue Analysis for Parkinsonian Gait Recognition.* Biomedical Engineering Journal, 2013. **7**: p. 1-8.

180. Pintea, S.L., et al., *Hand-tremor frequency estimation in videos.* Proceedings of the European Conference on Computer Vision (ECCV) 2018, 2018.

181. Lones, M.A., et al., *Characterising neurological time series data using biologically motivated networks of coupled discrete maps.* BioSystems, 2013. **112**(2): p. 94-101.

182. Lones, M.A., et al., *Evolving classifiers to recognize the movement characteristics of parkinson's disease patients.* IEEE Transactions on Evolutionary Computation, 2014. **18**(4): p. 559-576.

183. Noyce, A.J., et al., *Bradykinesia-akinesia incoordination test: Validating an online keyboard test of upper limb function.* PLoS ONE, 2014. **9**(4).

184. Andreotti, F., et al., *Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG.* Computing in Cardiology, 2017. **44**: p. 1-4.

185. Pembury Smith, M.Q.R. and G.D. Ruxton, *Effective use of the McNemar test.* Behavioral Ecology and Sociobiology, 2020. **74**(11): p. 133.

186. Hallett, M., *Overview of Human Tremor Physiology.* Movement Disorders, 1998. **13**(S3): p. 43-48.

187. Hallett, M., *Tremor: Pathophysiology.* Parkinsonism and Related Disorders, 2014. **20**(SUPPL.1): p. S118-S122.

188. Wade, P., M.A. Gresty, and L.J. Findley, *A Normative Study of Postural Tremor of the Hand.* Archives of Neurology, 1982. **39**(6): p. 358-362.

189. Beuter, A., et al., *Characterization of subclinical tremor in Parkinson's disease.* Movement Disorders, 2005. **20**(8): p. 945-950.

190. Alper, M.A., J. Goudreau, and M. Daniel, *Pose and Optical Flow Fusion (POFF) for accurate tremor detection and quantification.* Biocybernetics and Biomedical Engineering, 2020. **40**(1): p. 468-481.

191. Hemm-Ode, S., et al., *Intraoperative optical flow based tremor evaluation -a feasibility study.* Biomedizinische Technik, 2013. **58**(SUPPL. 1 TRACK-A): p. 1-2.

192. Uhríková, Z., et al., *Validation of a new tool for automatic assessment of tremor frequency from video recordings cka.* Journal of Neuroscience Methods, 2011. **198**: p. 110-113.

193. Williams, S., et al., *Computer vision of smartphone video has potential to detect functional tremor.* Journal of the Neurological Sciences, 2019. **401**: p. 27-28.

194. Liu, Y.J., et al., *A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition.* IEEE Transactions on Affective Computing, 2016. **7**(4): p. 299-310.

195. Dibia, V., *Building a Real-time Hand-Detector using Neural Networks (SSD) on Tensorflow: victordibia/handtracking*. 2018.

196. Elble, R.J., *The Essential Tremor Rating Assessment Scale.* Journal of Neurology and Neuromedicine, 2016. **1**(4): p. 34-38.

197. Roper, L.S., et al., *How to use the entrainment test in the diagnosis of functional tremor.* Practical Neurology, 2013. **13**(6): p. 396-398.

198. Mikalai, Z., et al., *Human body measurement with the iPhone 12 Pro LiDAR scanner.* AIP Conference Proceedings, 2022. **2430**: p. 090009-090009.

199. Urban, S., et al., *On the Issues of TrueDepth Sensor Data for Computer Vision Tasks Across Different iPad Generations.* 2022: p. 1-17.

200. Breitbarth, A., et al., *Measurement accuracy and dependence on external influences of the iPhone X TrueDepth sensor.* 2019. **1114407**(September 2019): p. 7-7.

201. Cooke, G., *Is it true that 'A careful history will lead to the diagnosis 80% of the time'?* Australian Family Physician, 2012. **41**(7): p. 534-534.

202. Hampton, J.R., et al., *Relative Contributions of History-taking, Physical Examination, and Laboratory Investigation to Diagnosis and Management of Medical Outpatients.* British Medical Journal, 1975. **2**(5969): p. 486-489.

203. Roshan, M. and A.P. Rao, *A study on relative contributions of the history, physical examination and investigations in making medical diagnosis.* The Journal of the Association of Physicians of India, 2000. **48**(8): p. 771-775.

204. Summerton, N., *The medical history as a diagnostic technology.* British Journal of General Practice, 2008. **58**(549): p. 273-276.

205. Deuschl, G., et al., *The clinical and electrophysiological investigation of tremor.* Clinical Neurophysiology, 2022. **136**: p. 93-129.

206. van der Stouwe, A.M.M., et al., *How typical are 'typical' tremor characteristics? Sensitivity and specificity of five tremor phenomena.* Parkinsonism and Related Disorders, 2016. **30**: p. 23-28.

207. Ericsson, L., et al., *Self-Supervised Representation Learning: Introduction, advances, and challenges.* IEEE Signal Processing Magazine, 2022. **39**(3): p. 42-62.