A data driven approach towards understanding deficits in olfactory and motor function and identifying applications of smartphone motor testing in prodromal and manifest Parkinson's

Christine Lo

The Queen's College

Supervisors:

Professor Michele Hu

Professor Maarten de Vos

Dr Siddharth Arora

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2020

A data driven approach towards understanding clinical tests of olfactory and motor function and identifying applications of smartphone motor testing in prodromal and manifest Parkinson's

Dr Christine Lo
The Queen's College

A thesis submitted for the degree of Doctor of Philosophy
Trinity Term 2020

Abstract

Background

Significant disease heterogeneity in Parkinson's means that identifying ways of applying observations made at a group level, to specific individuals, remains challenging. Standardised clinical assessments, critical for capturing the scientific evidence that informs clinical practice, are rarely used outside of research due to the time required for their administration.

Methods

Since 2010, the Oxford Discovery longitudinal cohort study has recruited over 1600 individuals with Parkinson's, Rapid Eye Movement Sleep Behaviour disorder (RBD) (a condition associated with a high risk of developing Parkinson's or another neurodegenerative disorder) and controls. Detailed clinical assessments, including clinical tests of olfactory and motor function and smartphone motor testing, were performed at 18-month intervals. Machine learning algorithms (chiefly random forests) were used to predict clinical scores and outcomes using clinical data or smartphone motor testing data alone.

Results

The use of three Sniffin' sticks (Anise, Liquorice, Banana) allowed the identification of individuals with a poor sense of smell (ordinarily requiring assessment with all 16 sticks) with excellent accuracy (area under the curve (AUCs) values of over 0.90 in development and validation cohorts). Summation of scores from 6 tasks of the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) motor examination: repetitive hand movements, finger tapping, pronation/supination, leg movements, constancy of rest tremor and bradykinesia, yielded an abbreviated score whose correlation coefficient with the total 18 task score was high at 0.91. Three approaches to using smartphone motor testing to quantify motor impairment were explored. Smartphone motor testing was also used to predict dopaminergic deficit in RBD with AUCs>0.73 and the new future onset of motor, cognitive and functional disability in Parkinson's with AUCs>0.75.

Conclusion

A data-driven approach can be used to promote understanding of existing clinical tests in prodromal and manifest Parkinson's, and to drive their refinement. Smartphone motor testing can be used to quantify motor impairment and provide individual estimates of risk; however, both are likely to benefit from further evaluation within the context of clinical trials and routine clinical practice.

Metrics

Pages:	422
Words:	42917
Figures:	132
Tables:	62

Figures, Tables, Table of Contents, List of Figures, List of Tables and References are excluded from metrics.

Table of Contents

List of abbreviations

AUC	Area Under the Curve
BRAIN	Bradykinesia-Akinesia Incoordination
CALM-PD	Comparison of the Agonist pramipexole vs. Levodopa on Motor complications in PD
CamPaIGN	Cambridgeshire Parkinson's Incidence from GP to Neurologist
CI	Confidence Interval
CoV	Coefficient of Variation
CSF	Cerebrospinal Fluid
CV	Cross Validation
DATATOP	Deprenyl And Tocopherol Antioxidative Therapy Of Parkinsonism
DaTscan	Dopamine Transporter imaging
DFA	Detrended Fluctuation Analysis
ECDF	Empirical Cumulative Distribution Function
FOG	Freezing Of Gait
GDNF	Glial cell line-Derived Neurotrophic Factor
GNE	Glottal to Noise Excitation
GPR	Gaussian Process Regression
GQ	Glottis Quotient
HADS	Hospital Anxiety and Depression Scale
HNR	Harmonics-to-Noise Ratio
HR	Hazard Ratio
ICC	Intraclass Correlation Coefficient
IQCODE	Informant Questionnaire on Cognitive Decline in the Elderly
KNN	K-Nearest Neighbours
LOSO	Leave One Subject Out
MAE	Mean Absolute Error

MDS	Movement Disorders Society
MDS-UPDRS	Movement Disorders Society-Unified Parkinson's Disease Rating Scale
MFCC	Mel Frequency Cepstral Coefficients
MIBG	iodine-131-metaiodobenzylguanidine
MLA	Machine Learning Algorithm
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
mPDS	Mobile Parkinson Disease Score
MSA	Multiple System Atrophy
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NPV	Negative Predictive Value
OPDC	Oxford Parkinson's Disease Centre
PARS	Parkinson Associated Risk Syndrome study
PCA	Principle Component Analysis
PD	Parkinson's Disease
PD-C	PD and Controls
PET	Positron Emission Tomography
PIGD	Postural Instability and Gait Difficulty
PNS	Peripheral Nervous System
PPE	Pitch Period Entropy
PPMI	Parkinson's Progression Markers Initiative
PPV	Positive Predictive Value
PSG	Polysomnography
PSP	Progressive Supranuclear Palsy
RBD	REM sleep Behaviour Disorder
RBD-C	RBD and Controls

RBDSQ	REM sleep Behaviour Disorder Screening Questionnaire
REM	Rapid Eye Movement
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
RPDE	Recurrence Period Density Entropy
SD	Standard Deviation
SPECT	Single Photon Emission Computed Tomography
SVM	Support Vector Machines
TD	Tremor Dominance
TIS	Temporal Irregularity Score
TKEO	Teager-Kaiser Energy Operator
TUG	Timed Up and Go
UK	United Kingdom
UPDRS	Unified Parkinson's Disease Rating Scale
UPSIT	University of Pennsylvania Smell Identification Test
VAS	Visual Analogue Scale
VFER	Vocal Fold Excitation Ratios
wMAPE	Weighted Mean Absolute Percentage Error

List of Figures

List of Tables

FOREWORD

Acknowledgements

This research would not have been possible, were it not for 1) the kindness and generosity of research participants, to whom we bear a duty to try to translate research findings back into routine clinical practice, or 2) the Oxford Discovery study team whose composition has varied since its inception in 2010 but whose participant centred approach has been great to observe and be a part of. I am particularly grateful to Francesca Bowring, Jessica Welch, Marie Crabbe, Katie Ahmed and Jane Rumbold in the Discovery office and to Tom Barber for taking the time to impart words of wisdom regarding thesis writing. Having benefited from the guidance of three supervisors, I am grateful to Michele Hu, for her strong support and patience; Maarten de Vos, for his invaluable technical insight and Siddharth Arora, who bore with me as a clinician and who helped to keep me on track in more ways than one.

Having made what felt like a difficult decision at the time, to embrace change and move to Oxford, I am exceedingly thankful to Michele for the opportunity I have had to grow academically, personally and spiritually. I am humbled by the continued support, understanding and acceptance I have received from Oliver Bandmann and Gary Dennis over the years and by the unconditional love and unswerving support of my family. Fully known and fully loved, above all, I am grateful to Him, in whom all things are possible.

"Teach us to number our days, that we may gain a heart of wisdom." – Psalm 90:12.

Funding

I am grateful to the National Institute for Health Research Oxford Biomedical Research Centre who funded my salary during the course of my DPhil. I am also grateful to The Queen's College for being elected to receive the Hastings Senior Scholarship and to the Nuffield Department of Clinical Neurosciences Staff Learning Scheme. The Oxford Discovery cohort study, from which, the majority of data in this thesis was obtained, was funded by Parkinson's UK.

Data and analyses

I have been involved in performing clinical assessments and biosampling of participants in the Discovery study since 2015, when I started recruiting and following up participants with Rapid Eye Movement sleep Behaviour Disorder (RBD) in Sheffield. Following my move to Oxford in 2017, I have been involved more extensively in the Discovery study, additionally seeing participants (both with RBD and Parkinson's) in the Thames Valley area and in Cambridge. I have also sought ethical approval, set up and managed the separate Wearables study (involving Oxford, Sheffield, Cambridge and Liverpool sleep centres), which is closely aligned to the Discovery study but whose focus is on testing wearable technology for the monitoring of sleep in RBD; comparing equipment to gold standard polysomnography and exploring its self-application at home.

All analyses described in this thesis were carried out by me except where explicitly stated (see sections 5.1 and 6.1). Other than Table 26 - Table 29 which have been reproduced with the permission of Dr Siddharth Arora, all work was authored by me. Pre-processing of raw smartphone recordings was performed by Dr Siddharth Arora in order to generate the smartphone features, which I utilised in analyses. DaTscans were

reported descriptively by Dr Kevin Bradley and reports were categorised by Dr Tom Barber.

List of publications

Elements of this thesis have been previously published or are in the process of submission.

CHAPTER 3:

- Parts of the introduction section to Chapter 3 have been contributed towards a review article on RBD which is to be submitted to The Lancet Neurology: Biomarkers of Phenoconversion in Isolated REM Sleep Behavior Disorder: A review by the international RBD study group
- Lo C, Arora A, Ben-Shlomo Y, Barber TR, Lawton MA, Klein JC, Kanavou S, Janzen A, Sittig E, Oertel WH, Grosset D, Hu MT. Olfactory testing in Parkinson's disease & REM behavior disorder; a machine learning approach. [submitted to Neurology]

CHAPTER 6:

- Lo C, Arora S, Baig F, Lawton MA, El Mouden C, Barber TR, Ruffmann C, Klein JC, Brown P, Ben-Shlomo Y, De Vos M, Hu MT. Predicting motor, cognitive & functional impairment in Parkinson's. Ann Clin Transl Neurol. 2019; 6(8):1498-1509. *Open access article under the terms of the Creative Commons Attribution License Attribution 4.0 International (CC BY 4.0).*

Other publications:

- Barber TR, Griffanti L, Bradley KM, McGowan DR, Lo C, Mackay CE, Hu MT, Klein JC. Nigrosome 1 imaging in REM sleep behaviour disorder and its association with dopaminergic decline. Ann Clin Transl Neurol. 2019; 10.1002/acn3.50962.
- Del Din S, Yarnall AJ, Barber TR, Lo C, Crabbe M, Rolinski M, Baig F, Hu MT, Rochester L. Continuous Real-World Gait Monitoring in Idiopathic REM Sleep Behavior Disorder. J Parkinsons Dis. 2019; doi: 10.3233/JPD-191773.
- Baig F, Kelly MJ, Lawton MA, Ruffmann C, Rolinski M, Klein JC, Barber T, Lo C, Ben-Shlomo Y, Okai D, Hu MT. Impulse control disorders in Parkinson and RBD: A longitudinal study of severity. Neurology. 2019; doi: 10.1212/WNL.0000000000007942.
- Kelly MJ, Lawton MA, Baig F, Ruffmann C, Barber TR, Lo C, Klein JC, Ben-Shlomo Y, Hu MT. Predictors of motor complications in early Parkinson's disease: A prospective cohort study. Mov Disord. 2019; doi: 10.1002/mds.27783.
- Postuma RB, Iranzo A, Hu M, Högl B, Boeve BF, Manni R, Oertel WH, Arnulf I, Ferini-Strambi L, Puligheddu M, Antelmi E, Cochen De Cock V, Arnaldi D, Mollenhauer B, Videnovic A, Sonka K, Jung KY, Kunz D, Dauvilliers Y, Provini F, Lewis SJ, Buskova J, Pavlova M, Heidbreder A, Montplaisir JY, Santamaria J, Barber TR, Stefani A, St Louis EK, Terzaghi M, Janzen A, Leu-Semenescu S, Plazzi G, Nobili F, Sixel-Doering F, Dusek P, Bes F, Cortelli P, Ehgoetz Martens K, Gagnon JF, Gaig C, Zucconi M, Trenkwalder C, Gan-Or Z, Lo C, Rolinski M, Mahlknecht P, Holzknecht E, Boeve AR, Teigen LN, Toscano G, Mayer G, Morbelli S, Dawson B, Pelletier A. Risk and predictors of dementia and

parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. Brain. 2019;42(3):744-759.

- Cooray N, Andreotti F, Lo C, Symmonds M, Hu MTM, De Vos M. Detection of REM sleep behaviour disorder by automated polysomnography analysis. Clin Neurophysiol. 2019; 130(4):505-514.
- Cooray N, Andreotti F, Lo C, Symmonds M, Hu MTM, De Vos M. Automating the Detection of REM Sleep Behaviour Disorder. Conf Proc IEEE Eng Med Biol Soc. 2018;1460-1463.
- Andreotti F, Phan H, Cooray N, Lo C, Hu MTM, De Vos M. Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks. Conf Proc IEEE Eng Med Biol Soc. 2018:171-174.
- Arora S, Baig F, Lo C, Barber TR, Lawton MA, Zhan A, Rolinski M, Ruffmann C, Klein JC, Rumbold J, Louvel A, Zaiwalla Z, Lennox G, Quinnell T, Dennis G, Wade-Martins R, Ben-Shlomo Y, Little MA, Hu MT. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. Neurology. 2018;91(16):e1528-e1538. *Open access article under the terms of the Creative Commons Attribution License Attribution 4.0 International (CC BY 4.0).*
- Barber TR, Griffanti L, Muhammed K, Drew DS, Bradley KM, McGowan DR, Crabbe M, Lo C, Mackay CE, Husain M, Hu MT, Klein JC. Apathy in rapid eye movement sleep behaviour disorder is associated with serotonin depletion in the dorsal raphe nucleus. Brain. 2018;141(10):2848-2854.

List of presentations

International

- 11th International REM sleep behaviour disorder meeting, Prague 12/10/17

    - "Using wearable technology to delineate and stratify RBD in the Oxford Discovery cohort"
- Movement Disorders Congress Hong Kong 9/10/18
    - "The use of smartphone task derived features to predict clinical scores in Parkinson's Disease"
    - Awarded Travel grant and selected to be part of the guided poster tour

National

- Parkinson's UK Research Conference 13/11/18
    - Smartphone prediction of clinical motor scores and future falls, freezing and disability in the Oxford Discovery cohort
- North of England Neurological Association meeting 24/11/18
    - Smartphone prediction of clinical motor scores and future falls and disability in the Oxford Discovery cohort

Regional

- Oxford Parkinson's Disease Centre (OPDC) 10/1/18
    - Delivered talk on "Using wearable technology to delineate and stratify RBD and PD in the Oxford Discovery cohort"
- Centre PD meeting 20/3/18
    - "Wearable technology for the diagnosis & monitoring of REM sleep behaviour disorder (RBD)"

List of patient and public involvement activities

- OPDC cohort day for research participants 4/9/17

    - “What is OPDC doing to improve your sleep?”
- OPDC cohort day for research participants 10/9/18
    - “Using smartphones to predict clinical measures and outcomes in Parkinson’s”
- OPDC cohort day for research participants 16/9/19
    - “Assessing sense of smell in RBD and PD”
- Smartphone application focus group 4/12/19
    - Co-led focus group to obtain feedback guiding the further development of the OPDC smartphone application

List of associated grants

- Co-applicant in awarded Cure Parkinson’s Trust grant for £211,207.68: Developing trial-ready wearable technology for Parkinson’s disease modification
- Co-PI in awarded Lab10x investment for £249,625: ParkAI: A mobile app with machine learning analysis for measuring and predicting clinical change in Parkinson’s

CHAPTER 1. Introduction

Note:

To assist cross referencing within this electronic thesis, all headers, figures and tables referenced within the text, use hypertext, which when clicked, allows users to jump directly to the referenced section.

1.1 Parkinson's disease

1.1.1 The burden of disease

Parkinson's disease (PD) is the second most common neurodegenerative disorder; a United Kingdom (UK) lifetime risk of 2.7% translates to 1 in every 37 people receiving a diagnosis of PD over the course of their lives.[1] Age poses the greatest risk factor for PD.[1, 2] The ageing population coupled with the rise in anticipated life expectancy of individuals living with PD, increasing industrialisation (and its by-products) and a reduction in smoking rates set the stage for a projected doubling in worldwide prevalence from 6.9 million in 2015 to 14.2 million by 2040.[3-5] The associated costs of PD, to the individual and to society are considerable. In a study that linked UK Clinical Practice Research Datalink and Hospital Episode Statistic data from 7271 individuals with PD and 7060 controls, the average overall total direct medical cost associated with PD was £5022 per year, over double the figure of £2001 for matched controls.[6] The indirect cost of living with PD, including social care costs and loss of income, additionally amount to an average of £16582 per person per year, according to a survey based study commissioned by the charity, Parkinson's UK.[7]

1.1.2 Clinical features

Originally the topic of “An essay on the shaking palsy” published in 1817, Dr James Parkinson described a syndrome characterised by involuntary tremors, a tendency towards a stooped posture and an “almost invincible propensity to run, when wishing only to walk”; the “supposed proximate cause” of the disease, in the “medulla spinalis…extending, as the disease proceeds, to the medulla oblongata”, remarkably reminiscent of theories of disease propagation in vogue 200 years later.[8] Indeed whilst estimates of dopaminergic neuron loss, within the substantia nigra pars compacta, at the time of motor symptom onset range from 30-80%, the pathological process starts much earlier.[9-11] Two theories, which need not be mutually exclusive, both involve the temporal spreading of prion-like alpha-synuclein pathology either 1) retrogradely, in the peripheral nervous system (PNS)-first hypothesis where rostral caudal spread first affects the autonomic PNS or 2) anterogradely, in the central nervous system-first hypothesis where the substantia nigra is affected before the autonomic PNS.[12-14]

Over the years, clinico-pathological correlation has led to the UK Parkinson’s Disease Society Brain Bank clinical diagnostic criteria being widely adopted. Sine qua non for the diagnosis of PD is the presence of a parkinsonian syndrome; bradykinesia (slowness of movement) accompanied by at least one of: rigidity, rest tremor or postural instability. Exclusion criteria include historical and examination features that would suggest an alternative aetiology, be it structural, traumatic, toxin- or medication-related, infective, vascular or another neurodegenerative disorder. Supportive criteria for a diagnosis of PD recognise the unilaterality of its onset, the persistence of asymmetry, the specificity of rest tremor as a feature, its responsiveness to levodopa and associated risk of dyskinesia (involuntary movements), and its slowly progressive clinical course (typically

>10 years).[15] Building on the Brain Bank criteria, the Movement Disorders Society (MDS) task force-derived criteria for PD, designed foremost for use in research, additionally incorporates olfactory loss, alongside evidence of cardiac sympathetic denervation on iodine-131-metaiodobenzylguanidine (MIBG) scintigraphy as supportive criteria for PD.[16]

Indeed, the potential for non-motor features of PD to negatively affect quality of life has been increasingly recognised. The pre-motor or prodromal phase of PD is characterised by olfactory dysfunction, sleep disturbance in the form of rapid eye movement (REM) sleep behaviour disorder (RBD) and excessive daytime sleepiness, mood disturbance typified by anxiety and depression, and constipation. Recent studies suggest that this prodromal phase can precede a clinical diagnosis of Parkinson's by up to twenty years.[17] In the early stages of motoric PD, fatigue, pain, apathy and mild cognitive impairment ensue. The onset of urinary symptoms and orthostatic hypotension tends to occur in the mid-late stages of the disease course and may coincide with motor fluctuations and dyskinesias. In the late stage of the disease, compounding disability from falls, postural instability, dysphagia and axial deformities, visual hallucinations and dementia become increasingly prevalent.[18, 19]

1.1.3 Diagnostic accuracy

Particularly early on in the disease course, PD can be challenging to differentiate from conditions that bear close homology including multiple system atrophy, progressive supranuclear palsy, corticobasal syndrome, benign tremor disorders, vascular parkinsonism and drug-induced parkinsonism. The current lack of a diagnostic test specific to PD means that a clinical diagnosis is considered the gold standard in life.[16]

However, following death, a definitive neuropathological diagnosis can be made in the presence of Lewy bodies; abnormal α-synuclein aggregated intraneuronal inclusions.[15, 20]

Clinical diagnostic accuracy rates are found to vary according to clinician experience and follow up duration. In a meta-analysis of studies that used a neuropathological diagnosis of PD as the standard of truth, the pooled diagnostic accuracy of non-experts was 73.8%, rising to 83.9% in movement disorder experts, having refined their diagnoses over follow up.[21] Whilst dopamine transporter imaging (DaTscan) has been approved by the United States Food and Drug Agency for the diagnosis of PD,[22-24] few studies have utilised neuropathological diagnosis as the ground truth when determining DaTscan diagnostic accuracy; those that have, suggest an accuracy roughly akin to that of clinicians.[24-26] The observed reduction in striatal binding has been suggested to correspond with the structural loss of nigral neurons rather than a functional reduction in dopamine transporter activity,[26] reflected in its primary use in differentiating PD from non-neurodegenerative causes of parkinsonism.[27]

1.1.4 Management

Non-pharmacological PD management strategies include ensuring access to PD specialist nurses, physiotherapy, high-impact exercise, occupational therapy, speech and language therapy and dietary advice. Exogenous dopamine replacement and potentiation of dopamine activity serve as the basis of pharmacotherapy for the motor symptoms of PD. Advanced therapies comprising deep brain stimulation, apomorphine pump delivery and continuous jejunal infusions of levodopa intestinal gel can be considered in individuals with motor symptoms responsive to medication but who have

developed motor complications. National Institute for Health and Care Excellence (NICE) guideline recommended pharmacological treatments for non-motor symptoms are symptom specific and include the use of clonazepam and melatonin for RBD.[28] Ultimately, the management of PD remains symptomatic and no treatment so far has been proven to slow down or stop disease progression.[29]

1.2 Rapid eye movement (REM) sleep behaviour disorder (RBD)

1.2.1 Definition

RBD is characterised by loss of the normal muscle atonia that occurs during the REM stage of sleep; the normal paralysis disrupted by dream enactment behaviour or vocalisations. RBD may initially occur in isolation, in association with other neurological disorders or secondary to medications including some commonly used antidepressants and anxiolytics.[30] References to RBD within this thesis (other than in sections 1.2.2 and 1.2.4), pertain to individuals with isolated RBD, who, at the time of diagnosis and initial recruitment, did not fulfil diagnostic criteria for a neurodegenerative disorder.

1.2.2 RBD in PD

RBD screening questionnaire (RBDSQ)[31] elicited rates of RBD, occurring within the context of established PD (38-57%),[32-36] are not too dissimilar from those benefiting from gold standard PSG confirmation (43-58%).[37-39] Compared to individuals with PD without RBD, individuals with concomitant PD and RBD have been found to be more likely to display an akinetic rigid subtype, report more frequent falls, be treated with higher levodopa dose equivalents, develop dyskinesia and dementia at accelerated rates, be more likely to have orthostatic hypotension and to more frequently be affected

by medication-induced psychosis and hallucinations; features indicative of a more aggressive phenotype.[39-42]

1.2.3 Isolated RBD

The community age-adjusted prevalence of isolated RBD in the over 50-60 year age group is estimated to be in the order of 1% in studies involving polysomnographic (PSG) confirmation.[43-45] Higher estimates of 6-8% are reported in studies solely reliant on questionnaire-facilitated symptomatic screening, carrying the risk of figure inflation due to affirmative answers from individuals who have non-REM parasomnias, seizures or forms of nocturnal movement disorders, unrelated to REM sleep.[46-48] Questionnaires that screen for symptoms of RBD include the RBDSQ,[31] the Hong Kong RBD questionnaire,[49] the Mayo sleep questionnaire[50] and the Innsbruck RBD inventory.[51] Their accuracy is dependent upon the population to which they are applied; the majority were originally derived from sleep clinic patient populations.[30] The Mayo sleep questionnaire, which has been validated in a community-based sample, relies on the availability of a bed partner for completion.[52]

In a study involving 203 individuals with RBD, of which the majority (84%) had a collateral history provided by a bed partner, 59% of individuals reported injury to themselves and 21% to their bed partner, as a result of nocturnal behaviours. The most common vocalisations reported were talking (96%) and screaming (90%); less frequently singing (15%) and barking (1%). Motor behaviours included punching (87%), kicking (82%) and falling out of bed (77%). Getting out of bed was reported in just under a quarter of individuals.[53] Despite the tendency of the nocturnal behaviours to be aggressive in nature, it is not uncommon for individuals to report a lack of awareness during the attacks; reported in 44% of individuals with isolated and 73% of individuals

with concomitant RBD and Lewy Body dementia.[53, 54] Ascertainment bias is likely to affect diagnosis; individuals self-presenting for evaluation of symptoms they or their bed partner have noticed, demonstrate motor events of higher severity than those detected on population screening.[55]

1.2.4 Management

Where intervention is required, management of RBD is symptomatic; environmental adjustments can help to reduce the risk of injury both to the individual with RBD and their bed partner; pharmacotherapy decisions are largely made on the basis of clinical experience or case series.[56, 57]

Clonazepam has benefited from being inexpensive and readily accessible, yet adverse effects are common; most frequently drowsiness, followed by unsteadiness and difficulty thinking.[57] The one randomised placebo-controlled trial involving clonazepam, recruited individuals with concomitant PD and probable RBD (lacking PSG confirmation) and failed to demonstrate superiority of clonazepam (N=20) at a dose of 0.5mg over placebo (N=20), utilising investigator-rated clinical global impression of improvement as the primary outcome measure.[58]

For a long time, randomised controlled trial evidence for melatonin in RBD came from a single study that under-recruited.[59, 60] More recently, two small double-blind, randomised controlled trials evaluated the effect of prolonged-release melatonin 1) at a dose of 4mg in individuals with concomitant RBD and PD (N=15) versus placebo (N=15) and 2) at doses of 2mg (N=7) or 6mg (N=6) versus placebo (N=9) in individuals with isolated RBD. Neither trial found a significant difference between melatonin and placebo groups

in their primary endpoints; self-reported or clinician-judgement-based measures of severity.[61, 62] It remains to be seen whether higher doses of melatonin will prove efficacious; in a survey study of individuals followed up within a sleep service, around half of individuals treated with melatonin required doses higher than 6mg in order to experience a beneficial effect.[57]

1.2.5 Clinical course and neurodegenerative risk

Individuals with isolated RBD demonstrate subtle evidence of motor impairment on clinical examination and semi-quantitative tests of motor function, that is insufficient to make a clinical diagnosis of Parkinson's. In contrast, non-motor deficits can equal and exceed those observed in PD; with depression, anxiety and apathy exhibiting particular prominence.[63]

Individuals with isolated RBD are at risk of phenoconversion. In a multicentre study involving 1280 individuals with PSG confirmed isolated RBD, the risk of conversion to an overt neurodegenerative disorder was quantified at 6.3% per year; 56.5% first developing parkinsonism (including 4.5% with Multiple System Atrophy (MSA)) and 43.5% first developing dementia.[64] The high predictive value of PSG-proven RBD for prodromal PD is translated into a high positive likelihood ratio of 130 within the MDS Research Criteria for Prodromal PD, more than three times that associated with an abnormal dopaminergic positron emission tomography (PET) / Single photon emission computed tomography (SPECT) scan and higher than any other clinical motor or non-motor marker.[65]

In a study that back-extrapolated the onset of motor and non-motor manifestations of 55 individuals with isolated RBD who went on to phenoconvert, olfactory dysfunction was the earliest sign of an evolving underlying neurodegenerative process, estimated to occur over 20 years before phenoconversion. Impaired colour vision, constipation and erectile dysfunction were predicted to occur over the ensuing decade before the onset of slight urinary dysfunction and subtle cognitive impairment 7-9 years pre-conversion. The onset of subtle motor symptoms was found to portend conversion by 7-11 years and preceded the detection of abnormalities on motor examination; bradykinesia (5 years-) followed by rigidity (3-4 years-) and then rest tremor (1-2 years-pre-conversion).[17]

There is growing interest in the recruitment of individuals with isolated RBD to neuroprotective treatment trials, aimed at slowing down or stopping conversion completely. Yet the lengthy interval over months to years during which phenoconversion may occur, relative to the typical clinical trial duration, makes identifying individuals at greatest risk of imminent conversion of key importance in determining trial feasibility and the likelihood of a neuroprotective effect being demonstrated.[66-68]

Clinical features have been used to stratify risk for clinical trials. Sample size estimates are reduced through the exclusion of individuals aged ≤55 in whom an antidepressant trigger is suspected and by restricting inclusion to those with either evidence of olfactory dysfunction (>25% sample size reduction) or at least 2 abnormalities on motor testing (out of: Timed up and go, Alternate tap test, Purdue Pegboard and the Unified Parkinson's Disease Rating Scale (UPDRS) III) (>40% sample size reduction).[66] Notably the 3-year phenoconversion risk hazard ratio (HR) associated with the presence

of at least 2 abnormalities on motor testing (HR 3.9) was greater than that associated with any other clinical or neurophysiological marker (including that of submentalis and anterior tibialis phasic activity (HR 2.75) reported in a separate study).[66, 69]

1.3 Overall aims

The overall aims of this theses are:

- To gain a greater understanding of existing clinical tests of olfaction and motor impairment in prodromal and established Parkinson's
- To identify applications of smartphone motor testing in prodromal and established Parkinson's

1.3.1 To gain a greater understanding of existing clinical tests of olfaction and motor impairment in prodromal and established Parkinson's

Time is a key barrier to the translation of research tools into routine clinical practice. Despite its role in disease stratification and the unreliability of self-reported smell status, olfaction is rarely, if ever, tested; established tests of olfaction can consume over a third of the average new appointment time in Neurology.[70] Similarly, motor signs play a fundamental role in the diagnosis of PD and the monitoring of response to treatment.[15] Yet the extent of examination and degree of documentation varies between and within clinicians. Evidence-guided standardisation of PD motor examinations has the potential to improve overall standards of clinical care and, with growing interest in big data in the National Health Service (NHS), to improve the quality of consolidated data available for analysis.[71]

CHAPTER 2 describes the Oxford Discovery cohort, the main source of data analysed in this thesis, and its longitudinal study design. Clinical assessments performed at each visit are detailed including the Sniffin' stick test of olfaction, the Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) motor examination and smartphone motor testing.

CHAPTER 3 explores the use of Sniffin' stick assessments of olfactory function in classifying individuals into PD, RBD and controls, and further seeks to derive an abbreviated smell test to stratify individuals, identifying those with a poor sense of smell according to age- and sex- specific normative data, through understanding the relative importance of each individual stick.

CHAPTER 4 focuses on the MDS-UPDRS motor examination in order to identify the individual examination items of greatest importance in quantifying overall motor function, determining disease severity and predicting PD subtype.

1.3.2 To identify applications of smartphone motor testing in prodromal and established Parkinson's

Despite its popularity as an outcome measure in clinical trials, limitations of the MDS-UPDRS motor examination include its recognised inter- and intra-rater variability, floor effects particularly early on in PD and its insensitivity to subtle change.[72, 73] With rising smartphone ownership worldwide, there is growing interest in its potential to provide objective assessments of motor impairment, in clinic and at home, at frequencies higher than would otherwise be possible through reliance on clinic attendance alone; the area of remote telemonitoring made particularly pertinent by the coronavirus pandemic that continues today, to which clinical and research services are having to adapt.[74, 75]

CHAPTER 5 explores three ways of using smartphone motor assessments to derive scores that can be used to quantify motor severity: 1) through the prediction of the results of existing standardised clinical assessments of overall PD motor function, gait, balance and dexterity, 2) through the creation of smartphone based digital motor scores,

independent of existing clinical assessments and 3) through the creation of a new composite clinical motor score which is then predicted by smartphone motor testing.

CHAPTER 6 investigates the use of smartphone motor testing, in the prediction of dopaminergic deficit on imaging (associated with phenoconversion risk) in RBD. Furthermore, the potential of smartphone motor testing to predict motor, cognitive and functional disability 18 months before its onset, in individuals with PD, is compared to existing prediction models that utilise clinical variables.

Findings are summarised at the start of the discussion section of each of the aforementioned chapters and again in ***CHAPTER 7***, where the following potential thesis contributions:

1. Equipping researchers and clinicians with tools for screening
2. Standardising and informing existing clinical care
3. Developing tools for objective motor assessments
4. Predicting dopaminergic deficit and future clinical change

and future avenues of investigation are reviewed.

CHAPTER 2. General Methods

2.1 The Discovery Cohort Study

2.1.1 Cohort composition

The longitudinal Oxford Discovery Cohort Study was started in 2010, initially recruiting individuals with early onset Parkinson's (within 3.5 years of their diagnosis) and age-matched controls (chiefly spouses and friends). Outpatient Neurology clinics in the Thames Valley area, serving a total population of 2.4 million, identified eligible individuals with Parkinson's, all of whom fulfilled the United Kingdom PD Brain Bank criteria for probable PD.[34, 76] In 2012 recruitment was extended to individuals with isolated RBD who were identified and seen at three Sleep Centres based at the John Radcliffe Hospital, Oxford; the Royal Papworth Hospital, Cambridge; and the Royal Hallamshire Hospital, Sheffield.[63] Polysomnographic confirmation of a clinically suspected diagnosis of RBD was a mandatory inclusion criterion for individuals with RBD, consistent with International Classification of Sleep Disorders criteria.[77] Additionally, individuals with RBD needed to not have already developed a neurodegenerative disorder at the time of their baseline visit; any such individuals were withdrawn from the study following their baseline assessment and their data excluded from analyses. Other obligatory inclusion criteria included age>18 years old, fluency in English and the absence of cognitive impairment/dementia that would preclude the provision of informed consent.

2.1.2 Ethical approval

Prospective approval for the study was granted by the South Central Oxford A Research Ethics Committee (reference numbers: 10/H0505/71 and 16/SC/0108) and recruitment was contingent upon the provision of written informed consent.

2.1.3 Longitudinal follow up

Following recruitment, participants were reviewed at 18-monthly intervals. At each visit participants with isolated RBD were assessed by a trained neurologist and research nurse/associate, using a structured protocol to look for features indicative of conversion. Participants with Parkinson's were assessed for atypical features and data pertaining to participants judged to have less than a 90% probability of Parkinson's at their latest clinic visit, as determined by trained researchers, were excluded from analyses. Participants unable to attend clinic in person e.g. due to intercurrent illness, relocation etc. were offered an abbreviated follow up appointment over the telephone (see Table 2 legend for more details). All control participants were seen in person for their baseline visit. 50 controls were invited to return for a single further in-person clinic visit. Routine follow up of controls was otherwise conducted over the telephone at 18-monthly intervals. Participants reaching one or more of the surrogate study end points were excused from further follow up (Table 1).

Table 1 | Surrogate study end point definitions

Endpoint	Definition
Severe dementia	MDS-UPDRS 1.1 score of 4 (Cognitive dysfunction precluding the ability to carry out normal activities and social interactions).
Severe motor disability	Hoehn and Yahr scale of ≥4 in an 'ON' state due to the progression of Parkinson's with the participant no longer feeling able to attend clinic due to their advanced physical state.
Dependency	Nursing or residential care home resident and no longer attending hospital appointments for Parkinson's.

Significant and severe disease comorbidity	Examples include advanced or disseminated cancer, persistent and severe renal failure warranting lifelong dialysis, advanced complications from end stage diabetes.

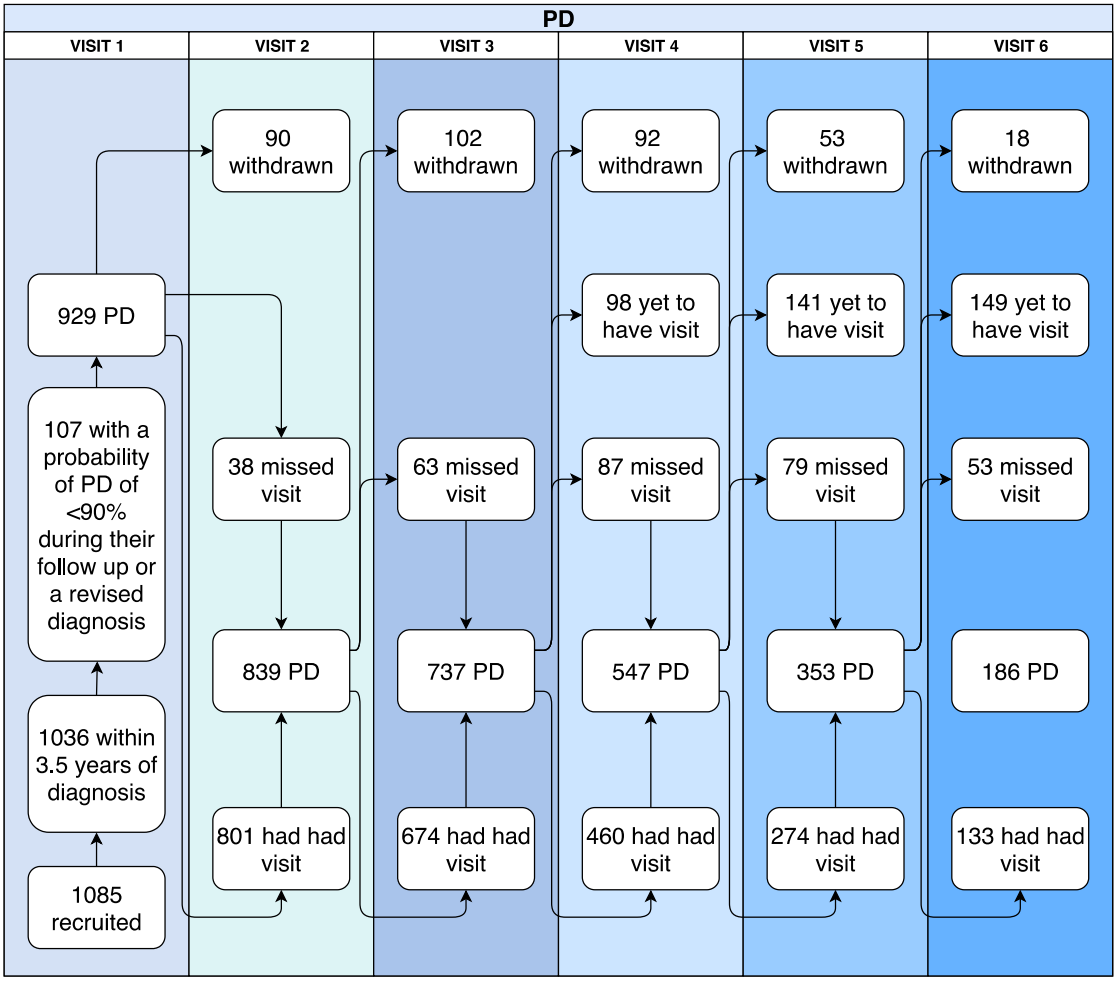
2.1.4 Cohort attrition

49 individuals with PD recruited into the study were found to have had a disease duration (from diagnosis) of more than 3.5 years and so were excluded (Figure 1). A further 62 had their initial PD diagnosis revised (19 Progressive Supranuclear Palsy (PSP), 2 PSP/Vascular parkinsonism, 5 Dementia with Lewy Bodies, 14 Multiple System Atrophy, 1 atypical Parkinson's not otherwise specified, 6 Vascular Parkinson's, 1 Cerebellar ataxia, 2 Essential tremor, 10 Dystonic tremor, 1 Atypical tremor, 1 Iatrogenic) (Figure 2) and an additional 45 individuals were deemed to have a probability of PD of <90% over the course of their follow up.

Of 258 individuals with RBD recruited into the study (Figure 3), 29 (11.2%) had converted as of 31st July 2019; 17 to PD, 8 to Dementia with Lewy Bodies or an unspecified dementia, 3 to Multiple System Atrophy and 1 to Pure Autonomic Failure.

Of 320 control participants recruited, as of 31st July 2019, none had developed idiopathic PD. 1 had developed vascular parkinsonism and had been withdrawn from further follow up (Figure 4).

Figure 1 | PD recruitment and retention



As assessed on 31st July 2019.

Figure 2 | Revised PD diagnoses

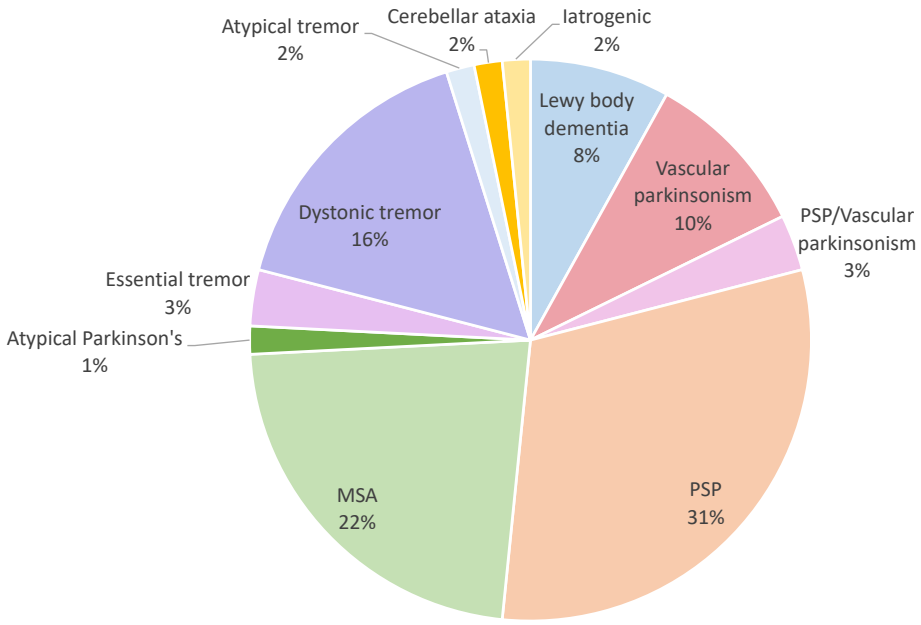
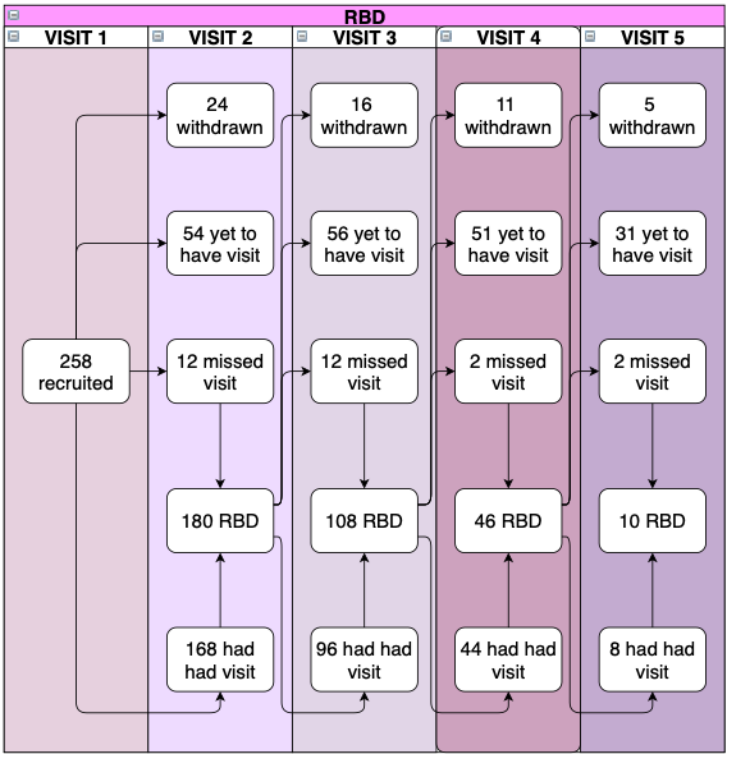
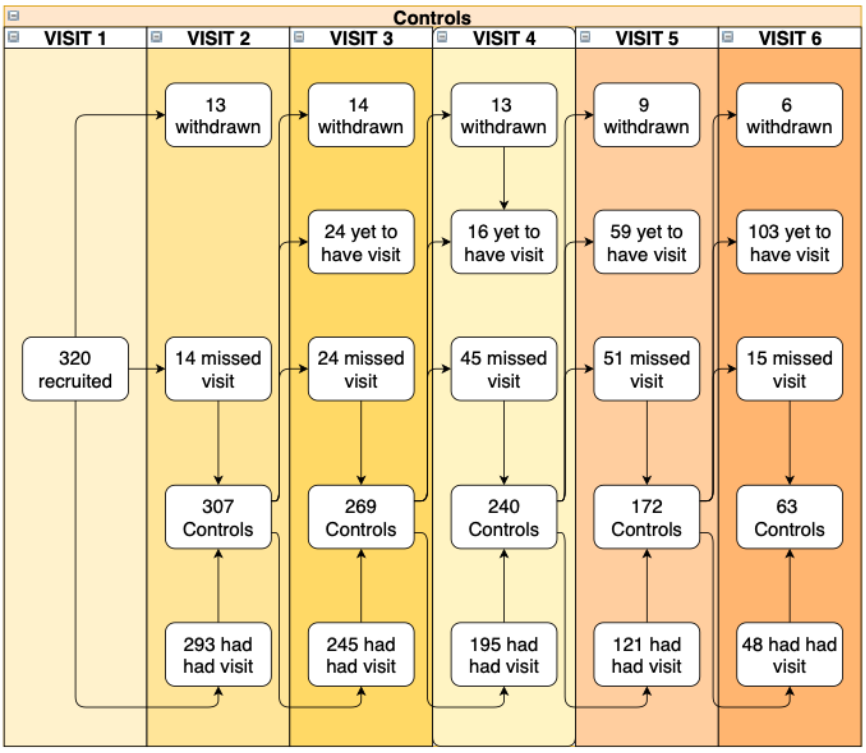


Figure 3 | RBD recruitment and retention



As assessed on 31st July 2019.

Figure 4 | Control recruitment and retention



As assessed on 31st July 2019.

2.1.5 Clinical assessments

Clinical assessments comprise a guided medical history, along with detailed non-motor, cognitive and motor assessments as summarised in Table 2.

Table 2 | Assessments performed as part of the Discovery study protocol.

	Control		RBD		PD		
Assessment	Visit 1	Visit 2 onwards[a]	Visit 1	Visit 2 onwards[b]	Visit 1	Visit 2&3	Visit 4 onwards[b]
Baseline assessments							
Demographics	X		X		X		
National Adult Reading Test	X		X		X		
Big Five Personality Inventory	X		X		X		
Medical history							
Past Medical History	X	X	X	X	X	X	X
MERQ-PD-B (Environmental exposure & lifestyle questionnaire)	X	X	X	X	X	X	X
PD features at diagnosis					X		
PD features at follow up						X	X
RBD Diagnosis dates			X				
Medication	X	X	X	X	X	X	X
Clinical global impression of change					X	X	X
Family history	X	X	X	X	X	X	X
Modified Schwab & England	X	X	X	X	X	X	X

Activities of Daily Living							
	Control		RBD		PD		
Assessment	Visit 1	Visit 2 onwards[a]	Visit 1	Visit 2 onwards[b]	Visit 1	Visit 2&3	Visit 4 onwards[b]
Social Background / Living arrangements	X	X	X	X	X	X	X
Non-motor assessments							
Sniffin' sticks	X		X	X	X	X	
MDS-UPDRS Part I	X	X	X	X	X	X	X
EQ-5D-3L (Health Questionnaire)	X	X	X	X	X	X	X
Constipation Questionnaire	X	X	X	X	X	X	X
Epworth Sleepiness Scale	X	X	X	X	X	X	X
RBD Screening questionnaire	X	X	X	X	X	X	X
Hospital Anxiety and Depression Scale (HADS)	X	X	X	X	X	X	X
Beck's Depression Inventory	X	X	X	X	X	X	X
Questionnaire for Impulsivity in Parkinson's	X	X	X	X	X	X	X
General Examination	X	X	X	X	X	X	X
Mini mental state examination (MMSE)	X		X	X	X	X	
Montreal cognitive assessment (MoCA)	X	X[c]	X	X	X	X	X
Phonemic Fluency	X	X	X	X	X	X	X
Semantic Fluency	X	X	X	X	X	X	X
Trail Making Test	X		X	X	X	X	X

	Control		RBD		PD		
Assessment	Visit 1	Visit 2 onwards[a]	Visit 1	Visit 2 onwards[b]	Visit 1	Visit 2&3	Visit 4 onwards[b]
Motor assessments							
Dopamine Challenge						Visit 2	
Freezing of gait questionnaire					X	X	X
Falls questionnaire	X	X	X	X	X	X	X
PD Screening Questionnaire	X	X	X	X			
MDS-UPDRS Part II motor Activities of Daily Living	X	X	X	X	X	X	X
MDS-UPDRS Part III	X		X	X	X	X	X
MDS-UPDRS Part IV					X	X	X
Unified Dyskinesia Rating Scale					X	X	X
26-item Parkinson Disease Dyskinesia Scale					X	X	X
Purdue Pegboard Test	X		X	X	X	X	X
Timed up and go	X		X	X	X	X	X
Flamingo test	X		X	X	X	X	X
Smartphone motor testing[c]	X	X	X	X	X	X	X
Collateral history							
Informant Questionnaire		X	X	X		X	X
Functional activities questionnaire[d]		X	X	X		X	X
Neuropsychiatric Inventory					X	X	X

[a]Controls were followed up over the telephone with the exception of 50 controls who were invited back for a single further in person clinic visit. [b]All participants were seen in person in clinic at baseline but for subsequent visits, for participants who were unable to attend clinic, a telephone appointment was offered; the MDS-UPDRS part III, MMSE, Trail Making Test, Unified Dyskinesia Rating Scale, Purdue Pegboard test, Timed up and go and the Flamingo test were omitted and an abbreviated version of the MoCA was administered over the phone. [c]Smartphone motor testing was introduced in 2014, four years after the study commenced recruitment to the PD and control arms so a minority of participants in those arms performed smartphone testing at their first study visit. [d]The functional activities questionnaire was introduced following a substantial amendment in 2018 by which time all controls and participants with PD had had their first visit.

The following clinical non-motor and motor assessments are of particular relevance to this thesis:

2.1.5.1 *Clinical non-motor assessments*

2.1.5.1.1 Sniffin' sticks assessment

First described by Hummel et al., the Sniffin' sticks test was touted as a cost effective, portable means of evaluating odour identification, odour discrimination and odour thresholds. The Sniffin' sticks test benefits from a large sample (n=9139) of normative data that has been used to define age and sex-specific smell classes.[78] Sniffin' sticks odour identification scores have also been equated to those of the longer standing, but more time consuming and costly University of Pennsylvania Smell Identification Test (UPSIT), involving single use scratch-and-sniff cards.[79]

In the Discovery study, sense of smell is assessed using the 16 Sniffin' sticks odour identification test (Burghart Instruments, Wedel, Germany). The felt-tip like pens are

presented to participants in turn who are asked to choose from one of four options (Table 3). Compared to testing odour discrimination and odour thresholds, odour identification is associated with superior test-retest reliability.[80] Smell is assessed at each clinic visit in participants with RBD and up to the third visit in participants with PD. Controls had their sense of smell assessed at baseline and a small proportion (n=40) had one further assessment during the course of their follow up. Before completing the odour identification test, participants are asked whether they have had a cold or flu within the preceding two weeks; those with an affirmative answer have their smell assessment rescheduled for a later date.

Table 3 | Forced choice options for the 16 Sniffin' sticks[80] odour identification test.

Stick	Options			
1	Orange	Strawberry	Blackberry	Pineapple
2	Smoke	Leather	Glue	Grass
3	Honey	Vanilla	Chocolate	Cinnamon
4	Chive	Fir	Peppermint	Onion
5	Coconut	Walnut	Banana	Cherry
6	Peach	Lemon	Apple	Grapefruit
7	Liquorice	Spearmint	Cherry	Cookies
8	Mustard	Menthol	Rubber	Turpentine
9	Onion	Garlic	Sauerkraut	Carrots
10	Cigarettes	Wine	Coffee	Smoke
11	Melon	Orange	Peach	Apple
12	Cloves	Cinnamon	Pepper	Mustard

13	Pear	Peach	Plum	Pineapple
14	Camomile	Rose	Raspberry	Cherry
15	Anise	Honey	Rum	Fir
16	Bread	Cheese	Fish	Ham

2.1.5.1.2 Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part I[81]

Providing a measure of non-motor symptom severity, Part I of the MDS-UPDRS involves the rating of 13 symptoms on a 5-point scale, ranging from normal (0 points) to severe (4 points). Non-motor symptoms assessed include cognitive impairment, hallucinations and psychosis, depressed mood, anxious mood, apathy, features of dopamine dysregulation syndrome, night time sleep problems, daytime sleepiness, pain and other sensations, urinary problems, constipation problems, light-headedness on standing and fatigue. Individuals are instructed to select the answer that best represents their average function over the preceding week and to rate the presence of symptoms independent of disease aetiology.

2.1.5.1.3 RBD Screening questionnaire (RBDSQ)[31]

Validated as a tool to screen for RBD, individuals are asked whether they agree or disagree with different statements relating to dream enactment behaviour. A cut-off of ≥5 was found to be associated with a sensitivity of 96% and specificity of 56% in distinguishing individuals with RBD from those with other sleep disorders in the original study; pooled sensitivity and specificity values, across 10 studies, were 91% (95% Confidence Interval (CI) 85%-95%) and 77% (95% CI 66%–85%) respectively.[31, 82]

2.1.5.1.4 Montreal Cognitive Assessment (MoCA)[83]

A 30-point test that assesses executive function, visuospatial abilities, language, short-term memory recall, phonemic fluency, verbal abstraction, attention, concentration, working memory and orientation. Scores are adjusted for education; an extra point being awarded to individuals with ≤12 years of education.

2.1.5.1.5 Modified Schwab and England activities of daily living scale[84]

Individuals are rated from 0% to 100%. 100% indicates complete independence and the absence of any functional difficulty. Each decile reduction signifies increased slowness and difficulty with activities of daily living. A score below 80% denotes loss of independence and a need for help. A score of 10% is given to individuals who are totally dependent and a score of 0%, to individuals who are bed ridden and have lost swallowing ability and bladder and bowel function.

2.1.5.1.6 EQ-5D-3L (Health Questionnaire)[85]

In the first part of the questionnaire, individuals are asked to rate their degree of symptomatology within five domains: mobility, self-care, usual activities, pain/discomfort and anxiety/depression, on a three-point scale, providing a measure of their overall health state. The second part of the questionnaire comprises a visual analogue scale (VAS), ranging from ‘Worst imaginable health state’ (equating to a score of 0) to ‘Best imaginable health state’ (equating to a score of 100), facilitating self-reported health ratings.

2.1.5.1.7 Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE)[86]

The 16-item IQCODE questionnaire is completed at each visit by the spouse, friend or relative of the research participant. A comparative assessment is made between the cognitive ease at which activities of daily living are performed by the research participant at the time of their visit, compared to 10 years ago. Performance is graded on a 5 point scale ranging from much improved to much worse, with higher total scores suggesting greater functional disability.

2.1.5.2 *Clinical motor assessments*

2.1.5.2.1 Timed up and go test (TUG)[87]

Individuals are seated in a chair with arms. The length of time taken to arise from the chair, walk 3m in a straight line to a marked line on the floor, turn and walk back to the chair and sit back down is measured. The task is performed three times and may be carried out using walking aids. The result taken is the best of the three attempts. The timed up and go test was added to the study protocol shortly after the study started; 22 participants with PD did not perform the test as part of their baseline assessments.

2.1.5.2.2 Flamingo test[63]

Individuals are asked to stand on one leg (the leg chosen, dependent on their preference) for a maximum of 30 seconds and the time achieved, recorded. The Flamingo test was added to the study protocol shortly after the study started; 22 participants with PD did not perform the test as part of their baseline assessments.

2.1.5.2.3 Purdue pegboard test[88]

Providing a measure of dexterity, individuals are given a total of 30 seconds to insert as many metal pegs into holes in a board 1) one at a time using their left hand 2) one at a time using their right hand 3) using both hands at the same time. A total score is calculated through the summation of the numbers of pegs inserted by each hand separately along with the numbers of pairs of pegs inserted by both hands simultaneously. A further assembly task is carried out over 60 seconds and involves the construction of what looks like cotton reels with both hands alternating in their application of 1) a metal peg, 2) a washer, 3) a collar 4) a washer. Throughout this thesis, analyses relating to the Purdue pegboard test are performed using the total score provided by each participant as opposed to the assembly task score.

2.1.5.2.4 MDS-UPDRS part II-III[81]

Parts II and III of the MDS-UPDRS assess self-reported and clinician-rated motor severity respectively. As with Part I of the MDS-UPDRS, each item is rated on a 5-point scale from 0 (normal) to 4 (severe). Part II focusses on the impact of motor disability on activities of daily living, including its impact on communication, salivation, eating, dressing, personal hygiene, handwriting, hobbies, turning in bed, tremor, getting out of bed, a car, or a deep chair, walking and balance and freezing. Part III, with a maximum total score of 132 points across 33 items involves a guided motor examination where examiners are instructed to “rate what you see”, without heed to causality. Examiners assess speech, facial expression, rigidity (stiffness of movement in the neck and four limbs), finger tapping, hand movements (opening and closing a fist), pronation/supination (turning the palm up and down), toe tapping, leg agility (stomping a foot on the floor), the ability to arise from a chair unaided, gait, freezing, postural stability

(where individuals are suddenly pulled backwards and the number of steps taken to stop themselves from falling counted), posture, global bradykinesia (slowness of movement), postural tremor (with arms outstretched), kinetic tremor (movement related) and rest tremor along with its constancy (the proportion of time rest tremor is present).

2.1.5.2.5 Hoehn and Yahr stage[89]

Intended to serve as a reproducible way of rating functional motor disability in Parkinson's, stages I and II distinguish unilateral and bilateral involvement, both with preserved balance. Stage III heralds the onset of impaired postural reflexes associated with mild to moderate disability. Stages IV and V indicate severe disability; the differentiating factor being the ability to walk or stand unassisted as opposed to being otherwise confined to a bed or wheelchair.

2.1.5.2.6 Freezing of gait (FOG) questionnaire[90]

A researcher-administered 6-item questionnaire. Like the MDS-UPDRS, symptoms are rated on a 5 point scale ranging from 0 to 4. The first two questions assess speed of walking and the functional impact of gait difficulties. The remaining four questions pertain to freezing frequency, its longest duration, as well as start and turning hesitation durations respectively.

2.1.5.2.7 Falls questionnaire

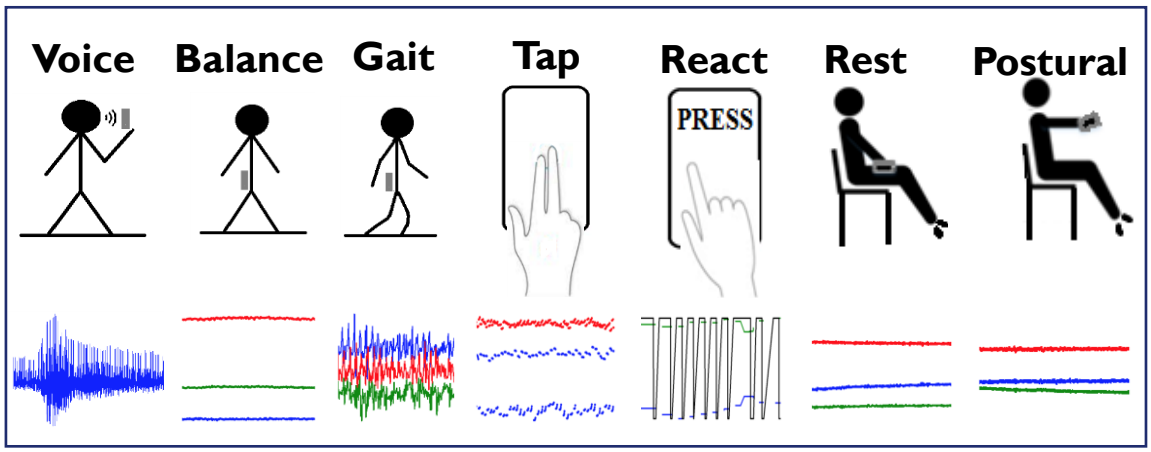
Participants self-report the number of falls they have had in the preceding 6 months.

2.1.5.2.8 Smartphone motor testing within the Discovery cohort

Smartphone testing was introduced to the Discovery cohort in 2014. Since then, individuals have performed tests in clinic and at home for up to 7 days. The Oxford Parkinson's Disease Centre (OPDC) configured Hopkin's PD smartphone test guides users through 7 tasks. Voice is assessed by asking participants to hold the phone to their ear, as if making a telephone call, before they take a deep breath in and say "aaah". With the smartphone positioned in their trouser pocket, balance is assessed whilst participants stand still. Similarly, gait is assessed with participants walking forward a distance of 20 yards in a straight line, before turning around and walking back to where they started from. To assess dexterity, finger tapping is assessed by individuals tapping the index and middle fingers alternately between two buttons on the smartphone screen. Non-cued reaction time is assessed by instructing participants to press on a button that appears on the screen as soon as they see it, hold their finger down whilst it is there and then release their finger as quickly as possible when it disappears, the button appearing and disappearing unpredictably over the course of the task. Rest tremor is assessed with the smartphone placed in the palm of the hand, with the hand at rest on a flat surface; on a table or on the arm of a chair, with their eyes closed as they count backwards from 100. A similar set up is used to assess postural tremor but with the arm outstretched in front of them, with their eyes open and without any concurrent counting (Figure 5).

Smartphone testing is carried out using consumer grade Android smartphones of a variety of makes and models, with participants being loaned a research phone with the smartphone application installed, to take home in order to do smartphone assessments, having done so first in clinic.

Figure 5 | The 7 smartphone tasks forming the smartphone test.



2.2 Analyses

2.2.1 Software

All data analyses were carried out using MATLAB® software (R2018a; Mathworks®, USA). draw.io version 12.6.6 (https://www.draw.io) (J Graph Ltd.) was used to generate flow charts in this thesis.

2.2.2 Statistical significance

Statistical significance was inferred from $p<0.05$ unless otherwise specified.

2.2.3 Random forests

Machine learning techniques were utilised extensively during analyses for this thesis. Random forests classification or regression ensemble models were trained; the choice dependent on whether the aim of the model was to distinguish groups or to predict a continuous variable.[91] Random forests encompass an ensemble learning technique combining multiple decision trees with bootstrap aggregation. Random forests may be considered to be one of the more simplistic machine learning approaches, especially

given recent advances in deep learning. However, they are advantageous in that they do not require large training samples and can provide feature (variable) rankings, giving insight into the predictor importance of each feature and thereby informing model refinement. Random forests have been used successfully in smartphone based studies to distinguish individuals with isolated RBD, PD and controls with high accuracies and have been used to predict longitudinal clinical outcomes using baseline cytokine measures, allowing the ranking of individual cytokines according their importance in predicting each clinical outcome.[92, 93] Though random forests were selected on the basis of model interpretability, comparisons were made with other supervised algorithms, an overview of which is provided in Table 4.

Table 4 | Overview of machine learning classification and regression models

Model type	Description
Linear Regression • Linear Regression • Linear Regression Interactions • Linear Regression Robust • Stepwise Linear Regression	Regression method which aims to find a linear relationship between a continuous dependent variable and one or more independent predictor variables by fitting a line that minimises the sum of squared errors or absolute errors
Logistic Regression	Classification method which aims to model the probability of an output/class using a logistic model, whereby the predictor variables can be both categorical or continuous
Simple decision trees • Fine Tree • Medium Tree • Coarse Tree	Applicable to both classification and regression problems, this modelling aims to identify the mapping of predictor variables (represented in the branches) on to the dependent variable (represented in the leaf node)
Discriminant Analysis • Linear Discriminant • Quadratic Discriminant • Subspace	Classification method which assumes the Gaussian distribution of predictor variables and aims to find a combination of predictor variables that separates two or more classes
Support vector machines (SVM) • Linear • Quadratic	Applicable to classification or regression problems, this method aims to identify a hyperplane (multi-dimensional plane) that

• Cubic • Fine Gaussian • Medium Gaussian • Coarse Gaussian	provides the greatest separation between data points of different classes, whilst allowing for non-linear classification
Gaussian Processes • Squared Exponential • Matern 5/2 • Exponential • Rational Quadratic	Applicable to classification or regression problems, this method aims to infer a distribution over functions and can therefore be used for probabilistic modelling
K-Nearest Neighbours (KNN) • Fine • Medium • Coarse • Cosine • Cubic • Weighted	Applicable to classification or regression problems, this method aims to estimate the target variable based on the proximity of a new predictor variable to previously observed predictor variables (known as neighbours)
Ensembles • Boosted Trees • Bagged Trees (e.g. Random Forests Ensembles) • RUS Boosted Trees • Subspace discriminant • Subspace KNN	Applicable to classification or regression problems, the ensemble learning methods are based on the concept of majority voting and rely on combining predictions from multiple decision trees, discriminant or nearest neighbour learners

2.2.4 Cross validation (CV)

In keeping with language used by the machine learning community, we refer to predictor variables (or independent variables) as *features*. Models trained with a large number of features relative to the available number of observations are at risk of overfitting the data. Indeed, if the data used to train, for example a random forests model, are the same as that with which its accuracy is later tested, as may be the case in studies employing linear regression, falsely elevated accuracies of 100% may be obtained.[94, 95] CV was therefore used to mitigate the risk of model overfitting.[96] Mutually exclusive training and test sets were derived from the main dataset. Models were created using only the training data, allowing the assessment of their generalisability to similar unseen datasets.[94] Two common CV schemes were used; 10 fold CV and leave one subject out (LOSO) CV, the latter also known as subject-wise cross validation.

10-fold CV involved the random division of the dataset into 10 non-overlapping folds. Each fold in turn was set aside as the test set and later used to evaluate the prediction accuracy of the model trained using the remaining 9 folds. 10 repetitions of 10 fold CV, each comprising different random divisions of the dataset, were performed. In LOSO CV, each data sample from a given participant in turn formed the test set and was used to evaluate the prediction accuracy of the model trained using all the other data samples in the dataset.

2.2.5 Classification data imbalance

Where the aim of a model is to distinguish groups (classes), data imbalance between groups may be translated into falsely elevated prediction accuracies. If one group is significantly larger, the tendency of a trained model to predict the larger group may

purely reflect its higher prevalence within the data used to train it. Consequently, balancing of groups is important and was achieved through the random under-sampling of the majority group in a 1:1 ratio, prior to either 10 fold or LOSO CV; in both cases the process was repeated for a total of 10 repetitions, with each repetition comprising another randomly balanced set of data. Under-sampling of the majority class was chosen as the method of balancing instead of over-sampling of the minority class so as to avoid the somewhat artificial replication of multiple sets of identical data from the same individual.[97]

2.2.6 Feature selection

Features differ in the degree to which they contribute to a given model. The removal of redundant features can not only improve computational time but in some cases improves model accuracy.[98] Crucially, the process of selecting a subset of the most salient features can provide useful insights into the modelling, thereby improving overall model inference. Where there are a large number of statistical features, it is likely that some may be collinear and associated with noise. Although multicollinearity can influence model inference, it does not reduce the overall predictive accuracy of the model. Across the ensemble of trees associated with random forests (we chose 500 trees) the predictor importance for each feature was derived using Gini's diversity index criteria for binary splitting. The average predictor importance of each feature, across all trained models was then calculated, allowing features to be ranked in order of importance.

2.2.7 Accuracy measures

To summarise the accuracy of multiple trained classification models associated with the prediction of a binary outcome, the probability of each data sample in the test set belonging to a given group was compared to its known group and area under the curve (AUC) values calculated. In evaluating the prediction accuracy associated with a single trained classification model, sensitivity, specificity, positive predictive (PPV) and negative predictive values (NPV) were calculated.

To evaluate model prediction accuracy where the outcome was numerical, the mean absolute error between the predicted and true values was calculated. As a comparator, in part to assess the effect of machine learning, the naïve mean absolute error was calculated by using the median of the training set as the prediction for each item in the test set.

CHAPTER 3. Olfaction in RBD and PD

3.1 Contextualisation

With the aim of deriving abbreviated tests of olfactory dysfunction that can be utilised clinically, mindful of time constraints, the work within this chapter focusses on exploring the use of smell tests as an adjunct in the diagnosis as well as in the stratification of individuals within disease groups. Some of the work in this chapter (pertaining to stratification according to sense of smell) has been submitted to the journal Neurology and is under review. Parts of the introduction section have also been contributed towards a review article on RBD which is to be submitted to The Lancet Neurology.

3.2 Introduction

3.2.1 Olfactory heterogeneity in PD

By the early stages of PD, up to 90% are affected by impaired olfaction, yet heterogeneity still exists.[99, 100] Relative sparing of olfactory function has been reported in individuals who are compound heterozygous for Parkin mutations compared to non-carriers and those carrying a single mutation.[101] Conversely, glucocerebrosidase mutation carriers demonstrate a greater frequency of hyposmia compared to those with sporadic PD.[102]

3.2.2 Impaired olfaction in PD prognostication

Individuals with early PD with evidence of olfactory dysfunction at baseline, demonstrate an accelerated trajectory of cognitive decline.[103-108] In a study involving 98 drug naïve individuals with PD, concurrent hyposmia (assessed using the Cross Cultural Smell Identification Test) and RBD (ascertained by the REM sleep Behaviour Disorder

Screening Questionnaire (RBDSQ))[31] was associated with lower Mini-Mental State Examination (MMSE) scores and a tendency towards the akinetic rigid PD phenotype.[108] Baseline hyposmia has been found to predict psychometric deficits in global cognition, verbal memory and word reading speed after a 7 year interval[107] as well as cognition-related functional deficits in activities of daily living, sufficient for a diagnosis of PD dementia (hazard ratio: 3.29 (95% CI 1.44-7.52) having controlled for age, gender, motor severity and mild cognitive impairment) after a 6 year median follow up.[103] In a study utilising Parkinson's Progression Markers Initiative (PPMI) study data from patients with newly diagnosed PD, in a between-patient comparison, for every 1 point reduction in University of Pennsylvania Smell Identification Test (UPSIT) score at baseline an additional 0.02 decline in Montreal Cognitive Assessment (MoCA) points per year was observed over a 3 year follow up period.[104] When combined with age, RBDSQ, cerebrospinal fluid (CSF) $A\beta_{42}$ and caudate signal intensity on DaTscan imaging, raw UPSIT scores additionally predicted cognitive impairment at two years with an AUC of 0.80 (compared to 0.68 when using age alone).[105]

3.2.3 Olfaction as a marker for prodromal PD

The interval between the onset of olfactory dysfunction and the onset of motor symptoms of PD may be as much as twenty years.[109, 110] 42 out of 2462 (1.7%) healthy individuals, followed up over a mean period of 9.8 years as part of the Health, Aging and Body Composition study, developed PD. Individuals in the lowest smell tertile had a hazard ratio of 4.8 (95% CI 2.0-11.2) of incident PD compared to 1.3 (95% CI 0.5-3.6) in the highest tertile; male sex and white race demonstrating higher degrees of association on subgroup analyses.[109] Focussing on an at-risk cohort of individuals with idiopathic smell loss, the rate of incident PD over a similar follow up period (mean 8.1 years) was

9.8%, lending weight to the case for stratifying individuals according to sense of smell.[110]

3.2.4 Concurrence of olfactory dysfunction with DaTscan imaging

Olfactory dysfunction predicts evidence of dopaminergic deficit on imaging.[111, 112] In a study involving 30 patients evaluated for drug-induced parkinsonism with DaTscan imaging, 2/23 with a normal scan had an UPSIT score below the 10th age- and sex-specific centile compared to 6/7 with an abnormal scan; an age- and sex-adjusted odds ratio of 63 (95% CI 4.8-820).[111] Differences in the scores associated with subjectively reported changes in taste or sense of smell have also been noted between individuals with early PD with and without imaging evidence of dopaminergic deficit.[112] As part of a two-tiered screening strategy aimed at identifying individuals at risk of incident PD in the Parkinson Associated Risk Syndrome (PARS) study, hyposmia combined with dopaminergic transporter deficit predicted conversion within a 4-year period with a sensitivity, specificity, PPV and NPV of 74%, 97%, 67% and 98% respectively, highlighting the interesting potential of combining predictive markers.[113]

3.2.5 Olfaction in RBD

Higher rates of subjectively reported and objectively tested impaired olfaction are apparent in individuals with RBD compared to controls.[63, 114, 115] Of specific relevance to isolated RBD, abnormal olfaction has additionally garnered credence as a predictor of disease conversion; its hazard ratio of 2.62, based on pooled multi-centre data from over 600 individuals, exceeding that of all other non-motor markers;[64] its use alongside an age cut-off of ≥55 suggested to improve the identification of individuals at risk of imminent conversion, for inclusion in future neuroprotective trials.[66]

3.2.6 Clinical tests of olfaction

Discrepancies between rates of subjectively reported hyposmia and rates identified on objective testing of upwards of 20% underline the importance of formal olfactory testing.[116, 117] The Sniffin' sticks test[80] and the UPSIT[118] are two popular tests of olfaction; the former benefits from relative cost effectiveness as the felt-tip style pens can be presented to multiple individuals whereas the latter necessitates the use of a once-only set of scratch cards.

Despite the recognition of olfactory loss in manifest and prodromal PD, olfaction is seldom tested in clinical practice, in part due to the time associated with testing (over a third of the average new appointment time in Neurology).[70] To date, most studies have been centred around identifying cut-offs that can be used to distinguish disease groups from controls on the basis of performance on odour identification testing alone;[119-121] abbreviated tests of olfaction typically being derived with the same aim.[122-124] Hummel et al. and Mueller et al. each described the use of a subset of Sniffin' sticks for the stratification of individuals according to their sense of smell; their studies involving healthy controls and those presenting to the Ear, Nose and Throat department respectively.[125, 126] To our knowledge, the aforementioned abbreviated Sniffin' stick tests have yet to be applied to an RBD or PD population and an abbreviated test has yet to be created specifically for stratification of PD by smell.

3.3 Aims

The work in this chapter on olfaction in RBD and PD centres around two main aims:

3.3.1 Classification according to sense of smell

Group differences in sense of smell are delineated before absolute cut-offs and machine learning approaches are applied to distinguish PD, RBD and controls.

3.3.2 Stratification according to sense of smell

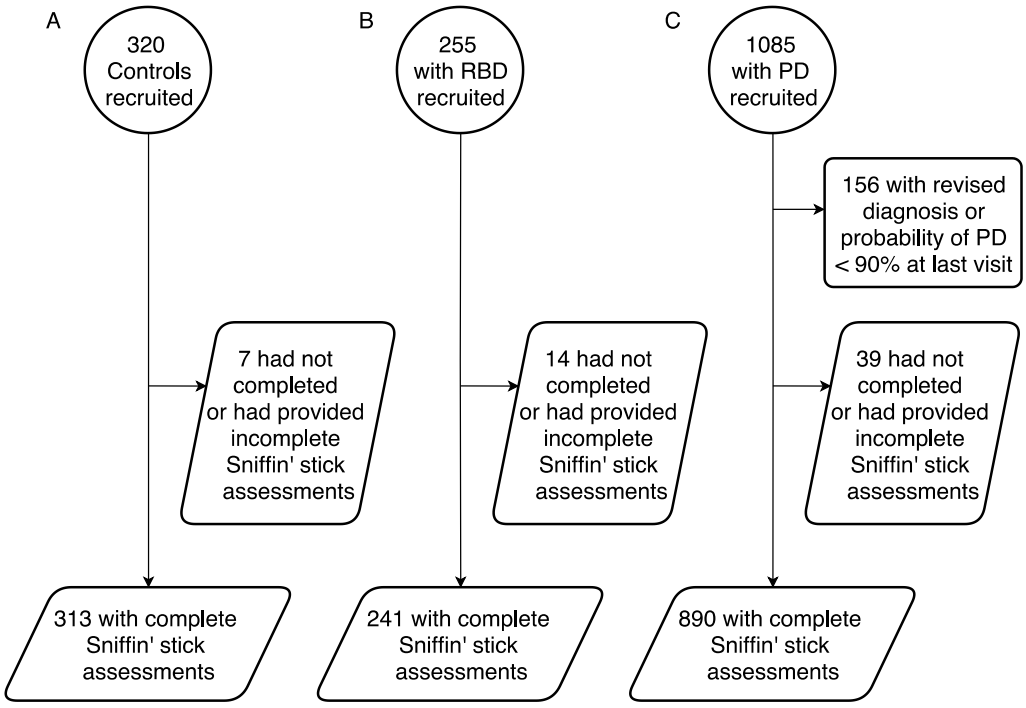
Individuals are categorised into those with a poor and good sense of smell. Differences in the rates of correct identification of individual smells, between smell classes within disease groups, are determined. An abbreviated test for the detection of poor smell is derived and validated using multiple independent datasets.

3.4 Methods

3.4.1 Participants

Sniffin' stick odour identification assessments were performed by participants in the Oxford Discovery study (described in the preceding General Methods chapter).[63, 100] Following the exclusion of incomplete smell assessments from 2% of Controls, 5% of individuals with RBD and 4% of individuals with PD, baseline data encompassed smell assessments from 313 Controls, 241 participants with isolated RBD and 890 participants with Parkinson's (Figure 6).

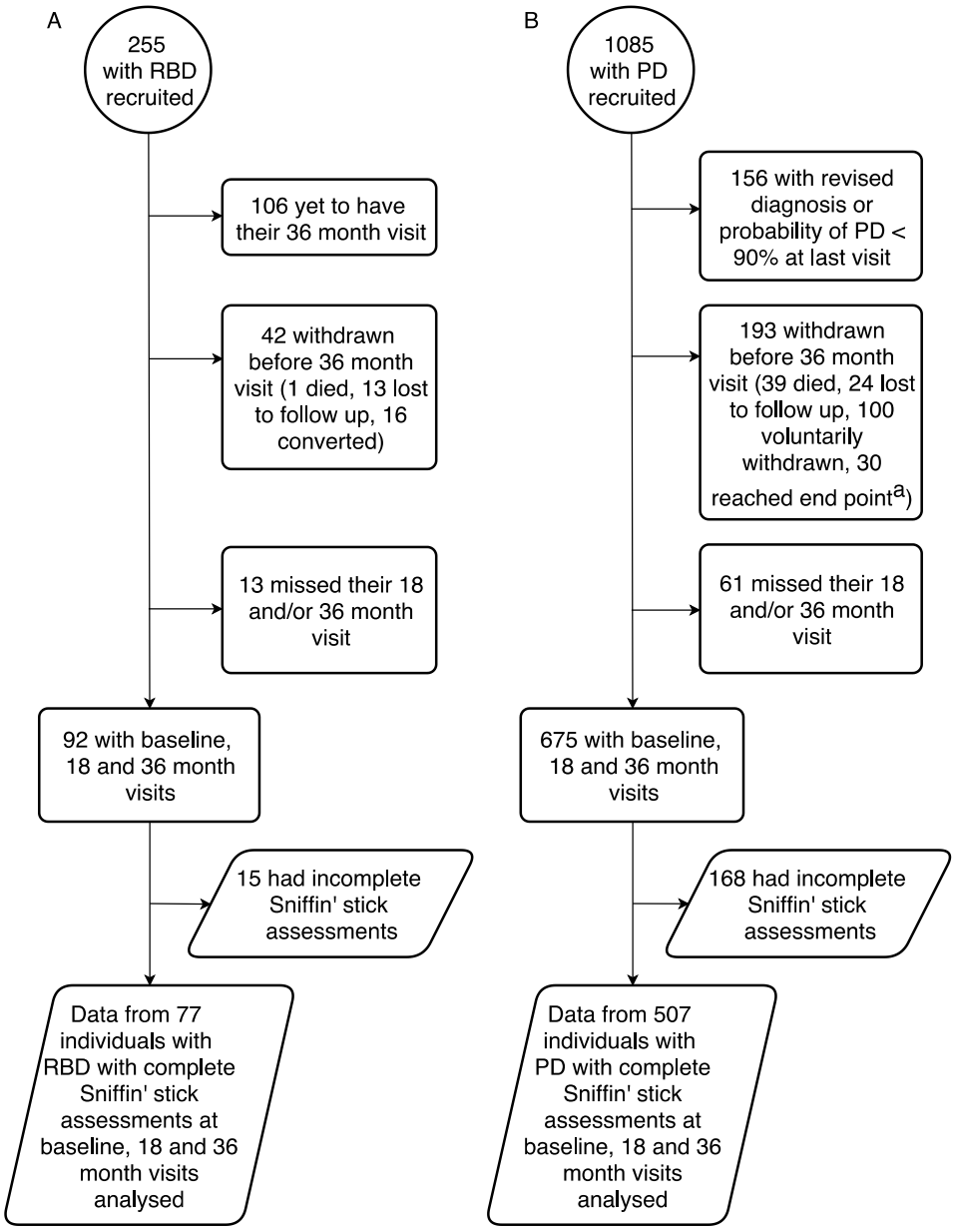
Figure 6 | Flowcharts illustrating baseline Sniffin' stick data provided by controls and participants with RBD and PD.



Participants in the Discovery study were seen at 18 monthly intervals. At the point of data download (28/5/19), 47 Controls had been invited back for a further in-person clinic visit (as opposed to typical follow up visits carried out via telephone) of which 40 had completed a further Sniffin' stick assessment, 92 participants with RBD had attended

their first 3 clinic visits of which 77 had contributed 3 complete sets of smell data and 675 participants with PD had had attended their first 3 clinic visits of which 507 had contributed 3 complete sets of smell data (Figure 7).

Figure 7 | Flow charts illustrating participants with RBD and PD who provided complete sets of Sniffin' stick data at baseline, 18 month and 36 month Discovery clinic visits.



[a]Surrogate study end point definitions as detailed in Table 1.

Independent RBD and PD datasets were provided by an RBD cohort based at the University of Marburg, Germany and the UK based Tracking Parkinson's cohort respectively. All 37 participants in the Marburg RBD cohort had had their clinically suspected RBD diagnosis confirmed on polysomnography. Whilst recruiting from distinct geographical areas, the Oxford Discovery and Tracking Parkinson's studies share similarities in their recruitment of individuals fulfilling the United Kingdom PD Brain Bank criteria for probable PD, within 3.5 years of diagnosis.[76, 127] The two cohorts have served interchangeably as independent validation cohorts for each other's findings.[127, 128] The UPSIT was the smell assessment of choice in the Tracking study; Sniffin' sticks assessments contributed by 452 participants were performed when there were difficulties in sourcing the UPSIT cards, with smell assessments being carried out up to 6 months from their baseline visit (at which time all other assessments were performed).

3.4.2 Clinical Assessments

Within the Discovery study, sense of smell was assessed through the administration of the 16 Sniffin' sticks odour identification test (Burghart Instruments, Wedel, Germany) (see 2.1.5.1.1 for further details). The test involves a forced choice paradigm whereby one of four options is selected by the participant following the presentation of each of the 16 felt-tip pens (Table 3).[78, 80] Sniffin' sticks were stored at room temperature, out of direct sunlight, in accordance with manufacturer instructions, and their replacement directed by the best before date displayed on each stick.

3.4.3 Statistical analyses

3.4.3.1 *CLASSIFICATION: Distinguishing individuals with PD, RBD and healthy controls using sense of smell*

3.4.3.1.1 Assessing group differences

Using baseline data, differences between groups in the total number of correctly identified Sniffin' sticks were assessed using paired t tests. Different cut-off thresholds were applied to the total Sniffin' stick score to distinguish PD and RBD from controls and each other and the AUC values for each comparison calculated. The optimal operating point of each Receiver Operating Characteristic (ROC) curve was identified and the associated cut-off threshold applied to datasets formed through the combination of external independent PD and RBD datasets, from the Tracking and Marburg studies respectively, with longitudinal control data from the Discovery study, which was treated as independent of the baseline control data; henceforth referred to as the classification validation datasets.

3.4.3.1.2 Adopting a machine learning approach to distinguish disease groups

A machine learning approach (see General methods chapter) was then adopted in order to train models to distinguish PD from controls, RBD from controls and PD from RBD using baseline smell assessments and the individual answers provided for each of the 16 Sniffin' sticks. Different machine learning approaches were compared using entire (unbalanced) datasets (RBD/control, PD/control and RBD/PD) where the raw answers provided for each of the 16 Sniffin' sticks were used alongside 10 fold cross validation to distinguish groups.

Data were then balanced, before the training of random forests and LOSO CV for a total of 10 repetitions, each involving a different randomly balanced set of data (see 2.2.4 for further details). Overall model accuracy was assessed through the calculation of AUC values and confirmed through the application of a single randomly selected trained model to each of the respective classification validation datasets. After establishing the individual predictor importance (for each Sniffin' stick), models were retrained using an incremental number of sticks, in descending order of importance. The accuracy of models trained was assessed using AUC values (whereby the data used to assess model accuracy had been excluded from the training process) and an abbreviated number of Sniffin' sticks for each pairwise comparison was selected for ongoing comparison. A single randomly selected abbreviated model for each comparison was selected and its accuracy assessed using the classification validation datasets described earlier.

3.4.3.1.3 Identifying absolute cut-offs for comparison

In order to ascertain the value of machine learning, all possible unique Sniffin' stick combinations, for a number of sticks equivalent to the abbreviated number selected for each pairwise comparison, were identified. Differing absolute cut-off values were then applied to the number of correctly identified sticks and AUC values were calculated. The Sniffin' stick combination associated with the highest AUC value for each comparison was selected and evaluated using the classification validation datasets.

3.4.3.2 *STRATIFICATION: Delineating sense of smell in PD and RBD*

3.4.3.2.1 Defining smell classes

In the largest Sniffin' stick study to date, Oleszkiewicz et al. categorised individuals into those with hyposmia, functional anosmia, normosmia or a super sense of smell using normative data from 9139 healthy individuals.[78] Their age and sex-specific thresholds were applied to the number of Sniffin' sticks correctly identified by participants with PD, RBD and healthy controls from the Discovery, Tracking and Marburg studies. Individuals who correctly identified 8 sticks or fewer were labelled as having functional anosmia; those not already classified as having functional anosmia, but with total Sniffin' stick scores below the 10th percentile for their age and sex, were classified as having hyposmia (Table 5).[78] Similarly, those with scores above the 90th centile were classified as super-smellers and the remaining individuals were classified as normosmic.[129] The hyposmia and functional anosmia groups were merged to form a poor smell group; normosmia and super-smellers were combined to form a good smell group and the rates of correct identification of each Sniffin' stick compared.

Table 5 | 10th and 90th centiles of odour identification total Sniffin' stick scores by age and sex

Age		**21 to 30**		**31 to 40**		**41 to 50**		**51 to 60**		**61 to 70**		**71 to 80**		**≥81**	
Sex		F	M	F	M	F	M	F	M	F	M	F	M	F	M
Centile	**10**	11	11	12	12	11	11	11	10	10	9	7	7	3.8	5
	90	16	16	15	15	15	15	15	15	15	15	14	14	13	13.3

Adapted from Oleszkiewicz et al. Updated Sniffin' Sticks normative data based on an extended sample of 9139 subjects. Eur Arch Otorhinolaryngol 2019;276:719-728 distributed under the terms of the Creative Commons Attribution 4.0 International License. F: female; M: male.

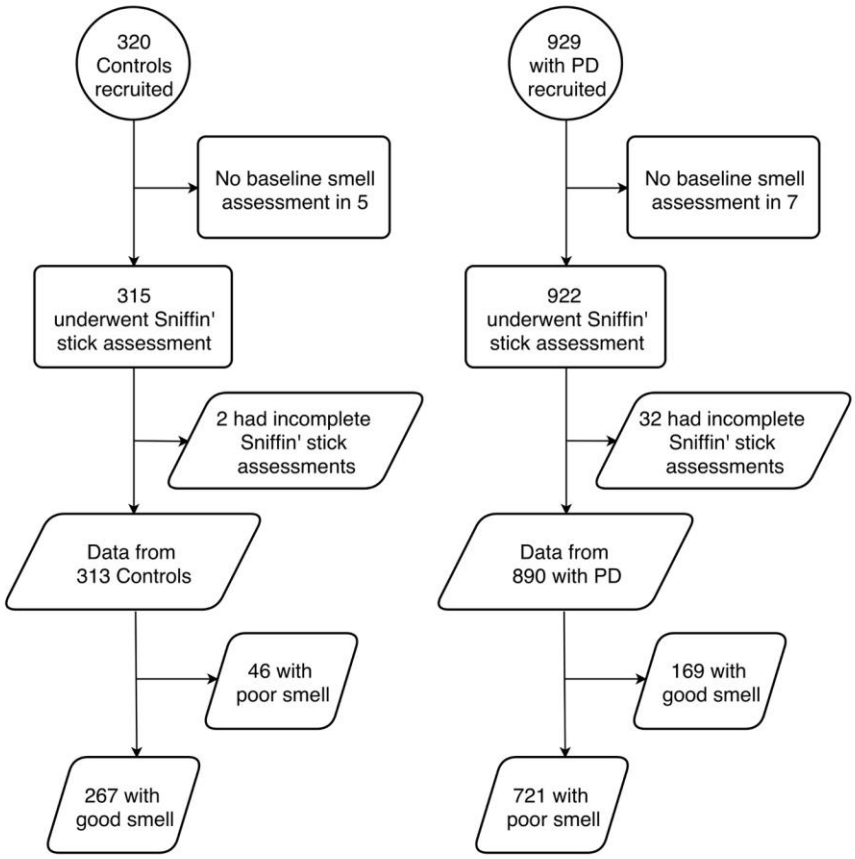
Data from 77 participants with RBD and 507 participants with PD who had had their sense of smell assessed at their baseline, 18 month and 36 month visits were used to determine smell class (i.e. hyposmia / functional anosmia / normosmia or super smell). Differences in smell class and total Sniffin' stick score between the baseline and 36 month visits were then compared to determine the stability in smell class / Sniffin' stick score over time.

3.4.3.2.2 Training machine learning algorithms to detect PD specific poor smell

With the aim of stratifying individuals according to the presence of a PD specific poor sense of smell, baseline Discovery data from individuals with PD and poor smell (n=721) and data from controls with good smell (n=267), representing the extremes of a smell continuum, were used to train models. Different machine learning approaches were compared using the entire PD poor smell/control good smell (unbalanced) dataset and 10 fold cross validation.

Data were then balanced and used to train random forests (Figure 8). A LOSO CV scheme was employed and a total of 10 repetitions, each involving a different randomly balanced group, led to a total of 5340 trained models whose accuracy was assessed using AUC values. Ranking of sticks according to predictor importance allowed models to be retrained with a sequential number of sticks selected according to averaged predictor importance and an abbreviated test to be created.

Figure 8 | Flow charts illustrating the PD poor smell and Control good smell data used to train the PD-specific poor smell models



3.4.3.2.3 Evaluating model performance

In selecting the minimal number of sticks to comprise the abbreviated test of smell, it was necessary to consider the inverse relationship between the number of sticks used and the AUC values associated with the trained models. The increase in AUC following the addition of 2 sticks to a single stick was threefold that achieved through the addition of the remaining 13 sticks. 3 sticks were therefore chosen to comprise the abbreviated test of smell; in keeping with other previously described abbreviated smell tests. From the 5340 models trained using the same top 3 sticks, one was randomly chosen for ongoing evaluation, hereafter referred to as the Discovery 3 Sniffin' stick model. Its accuracy was assessed internally, using baseline smell assessments from individuals

with RBD in the Discovery study and externally, using the independent Tracking (Parkinson's) and Marburg (isolated RBD) cohorts.

3.4.3.2.4 Comparison with previously described abbreviated Sniffin' stick combinations

The Discovery 3 Sniffin' stick model, comprising the top 3 sticks identified through machine learning was compared with abbreviated Sniffin' stick combinations previously described (Table 6): A) for the detection of poor smell (not specific to PD): the 3 Q stick test[126] and the 5 Brief Sniffin' Stick test[125] reported by Hummel et al. and Mueller et al. respectively. B) for the separation of individuals with PD from controls as reported by Casjens et al.,[122] Boesveldt et al.[130] and Mahlknecht et al.[123]

Table 6 | Sniffin' stick combinations previously described for the detection of poor smell and for distinguishing individuals with PD from controls.

	Detecting poor smell			Distinguishing PD versus controls		
	Discovery	Q Stick[a]	Brief Sniffin'[b]	Boesveldt et al.[130]	Casjens et al.[122]	Mahlknecht et al.[123]
Total sticks	3	3	5	3	3	8
Orange			■			
Leather			■			
Cinnamon				■		■
Peppermint			■		■	■
Banana	■					■
Lemon						
Liquorice	■			■		■
Turpentine						
Garlic						
Coffee		■			■	■
Apple						
Clove		■				
Pineapple						■
Rose		■	■			■
Anise	■			■	■	■
Fish			■			

[a]Hummel et al.[126]; [b]Mueller et al.[125]

3.4.3.2.5 Comparison with all other possible 3 stick combinations

There are a total of 560 possible unique 3 stick combinations of the 16 Sniffin' sticks. To compare their accuracy at detecting poor smell, a single balanced group was formed using baseline data from an equal number of randomly selected individuals from the PD poor smell group and the control good smell group. One model was trained using data from the balanced group for each 3 stick combination in turn and the accuracy of each model, in distinguishing poor smell from good smell, assessed using an independent composite validation set formed by combining the Discovery isolated RBD, Marburg isolated RBD and Tracking PD datasets. The process was repeated for 9 further randomly balanced training groups resulting in a total of 5600 trained models. The AUC values associated with our 3 Sniffin' stick model were compared individually with those associated with all other possible 3 stick combinations using pairwise t tests.

3.4.3.2.6 Evaluation of the effect of training dataset composition

In the creation of a PD specific poor smell model, data from individuals with PD and poor smell and controls with good smell, considered to be at either extreme of the smell spectrum, were used to train models. To evaluate the effect of different training datasets, a single balanced dataset was created using baseline data from 1) individuals with PD and poor smell and controls with good smell 2) controls with good smell and with poor smell and 3) individuals with PD with poor smell and good smell. Having established the balanced datasets in each case, a random forests model (comprising 500 classification trees) was trained for each of the 560 possible unique 3 stick combinations and the distributions of their associated AUC values in distinguishing poor from good smell in RBD was assessed using pairwise Kolmogorov-Smirnov tests.

3.4.3.2.7 Creation of a staged RBD screening model

The accuracy in distinguishing individuals with RBD from controls through the application of the predictors age≥55, RBDSQ≥5 and Discovery 3 Sniffin' stick predicted PD poor smell, alone and in combination was assessed using baseline Discovery RBD and longitudinal control data.

3.5 Results

3.5.1 Demographics

Table 7 illustrates group characteristics, mean total Sniffin' score and smell class for controls and individuals with RBD and PD in the different cohorts. There was no evidence of a difference in the total Sniffin' stick score of individuals with PD tested at 0-3 months following the start of the Discovery study and those tested at 9-12 months (p=0.26) or 15-18 months (p=0.10), suggesting no decline in odour testing with use.

Within disease groups there was no difference in sex between cohorts. Between disease groups, an expected male preponderance was observed in PD compared to controls; a difference even more pronounced in RBD.

Consistent with the 6-month interval between recruitment and baseline smell assessments in the Tracking cohort, participants with Parkinson's were older and of longer disease duration than those in the Discovery study, where baseline smell assessments were performed at the time of study recruitment. Lower MDS-UPDRS motor examination scores were evident in the Tracking cohort but were missing in 10% of participants. Participants with RBD in the Marburg cohort had had their diagnosis for longer than those in Discovery but both were of a similar age and motor severity.

Table 7 | Baseline demographics of participants in the Discovery study alongside the independent Tracking and Marburg datasets.

		Control	PD			RBD		
		Discovery		Tracking	P[a]	Discovery	Marburg	P[a]
Baseline demographics								
N		313	890	452	-	241	37	-
Age		64.4 (9.8)	66.5 (9.6)	68.0 (9.0)	<0.01	64.6 (8.9)	67.3 (8.8)	0.97
Male sex		165 (53%)	569 (64%)	290 (64%)	0.93	211 (88%)	34 (87%)	0.78
Disease duration at assessment		-	1.2 (0.9)	1.9 (1.0)	<0.001	1.4 (1.8)	3.5 (2.9)	<0.001
MDS-UPDRS part 3		1.8 (2.5)[b]	26.4 (10.8)	23.4 (12.8)[b]	<0.001	5.1 (5.9)[c]	5.0 (2.2)[d]	0.95
Montreal Cognitive Assessment		26.7 (2.5)[e]	25.0 (3.3)[e]	25.4 (3.4)[e]	0.03	25.1 (2.9)[e]	27.5 (2.1)[e]	<0.001
Subjective poor smell		34 (11%)[f]	540 (61%)	-	-	139 (58%)[f]	20 (59%)[g]	0.66
Objective smell testing								
Mean total Sniffin' score (SD)		12.1 (2.3)	7.1 (2.9)	7.4 (3.0)	0.04	7.9 (3.2)	7.5 (2.7)	0.44
Poor smell	Functional anosmia	25 (8%)	623 (70%)	309 (68%)	0.16	147 (61%)	24 (65%)	0.35
	Hyposmia	21 (7%)	98 (11%)	48 (11%)		23 (10%)	6 (16%)	
Good smell	Normosmia	217 (69%)	161 (18%)	84 (19%)		64 (27%)	7 (19%)	

	Super smell	50 (16%)	8 (1%)	11 (2%)		7 (3%)	0	

SD: Standard deviation. [a]p value determined using a two-sample t-test or chi squared test. [b]2 Control participants and 7 RBD participants and 11 PD participants from the Discovery study had an incomplete MDS-UPDRS part 3 score. Descriptive statistics relating to the MDS-UPDRS part 3 score for the aforementioned groups were therefore calculated across those with complete scores. [c]In the Tracking cohort MDS-UPDRS part 3 scores were assessed up to 6 months before the Sniffin' sticks test was performed. 45 participants did not have a motor examination within the 6 month window. [d]Motor impairment was assessed in the Marburg cohort using the original version of the UPDRS (as opposed to the MDS-UPDRS used in Discovery and Tracking). Scores were converted using the formula: (original UPDRS part 3 score x 1.2) + 2.3 which was developed to convert scores from individuals with PD at Hoehn & Yahr stage I&II.[131] Additionally 1 of the Marburg participants had their motor assessment 7 months after their smell assessment (as opposed to within a week of their assessment). [e]A total MoCA score was not available for 13 Controls, 37 participants with PD and 2 participants RBD in Discovery, 41 participants with PD in the Tracking cohort and 1 participant in the Marburg cohort. Summary statistics are calculated across those with available scores. Subjective smell status was missing in 1 participant in the Discovery control and RBD groups[f] and 2 participants in the Marburg RBD group.[g] It was not assessed in the Tracking study.

41% of controls, 84% of individuals with RBD and 89% of those with PD who self-reported poor smell had it confirmed on objective testing. Conversely, of those with objective evidence of impaired olfaction, normal subjective smell had been reported in 69% of controls, 32% of individuals with RBD and 34% of those with PD. Individuals with PD within the Tracking cohort had a higher mean total Sniffin' score, unadjusted for age or sex, compared to those in Discovery. However, following the application of age

and sex-specific normative data derived thresholds to the total Sniffin' score, there was no difference in smell class within disease groups.[78] Within the Discovery study, individuals with RBD and PD did not differ with respect to their frequency of poor smell on objective testing (p=0.67) but poor smell was more common in both groups compared to controls (p<0.001).

Of the 77 individuals with RBD who contributed Sniffin' stick data at their baseline, 18 month and 36 month visits, 61% had functional anosmia, 12% hyposmia, 25% normosmia and 3% super smell at their baseline visit. After 3 years of follow up 93% of those with poor smell at baseline continued to have poor smell and 38% of those with good smell at baseline had developed poor smell (Figure 10). There was no significant difference across smell classes (p=0.49) or in total Sniffin' stick score over the same period (p=0.63) (Figure 9).

Figure 9 | Total Sniffin' stick score in participants with RBD at baseline, 18 month (Visit 2) and 36 month (Visit 3) visits

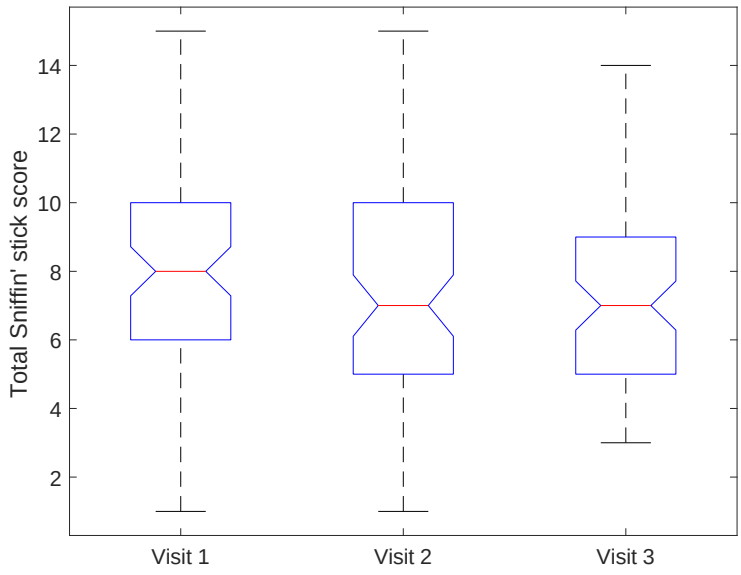
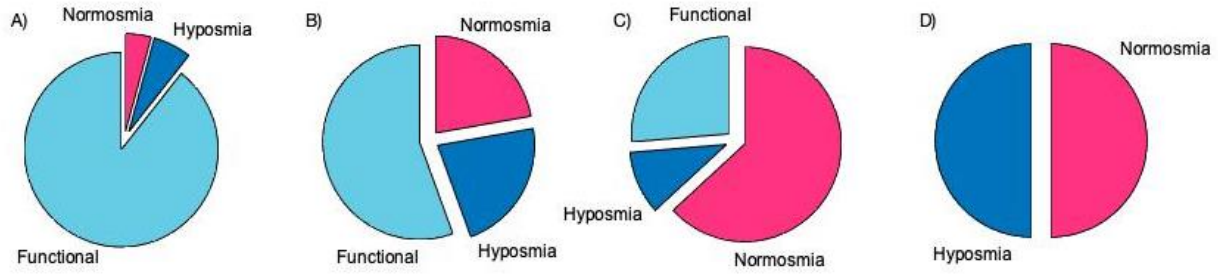


Figure 10 | Change in smell status between baseline and 36 month visits in individuals with RBD.



Each pie chart relates to individuals with a given smell status at baseline (Pie chart A = baseline functional anosmia, Pie chart B = baseline hyposmia, Pie chart C = baseline normosmia, Pie chart D = baseline super smell) and illustrates their smell status at their 36 month visit. i.e. of the individuals with super smell at baseline (Pie chart D), around half were classified as having hyposmia and the other half as having normosmia at their 36 month visit.

Of the 507 individuals with PD for whom there was Sniffin' stick data at their baseline, 18 month and 36 month visits, 69% had functional anosmia, 12% hyposmia, 17% normosmia and 1% super smell at their baseline visit. After 3 years of follow up 94% of those with poor smell at baseline continued to have poor smell and 52% of those with good smell at baseline developed poor smell (Figure 12). There was a significant difference across smell classes between baseline and 36 month visits ($p<0.0001$), and in total Sniffin' stick score over the same period ($p<0.0001$) (Figure 11).

Figure 11 | Total Sniffin' stick score in participants with PD at baseline, 18 month (Visit 2) and 36 month (Visit 3) visits

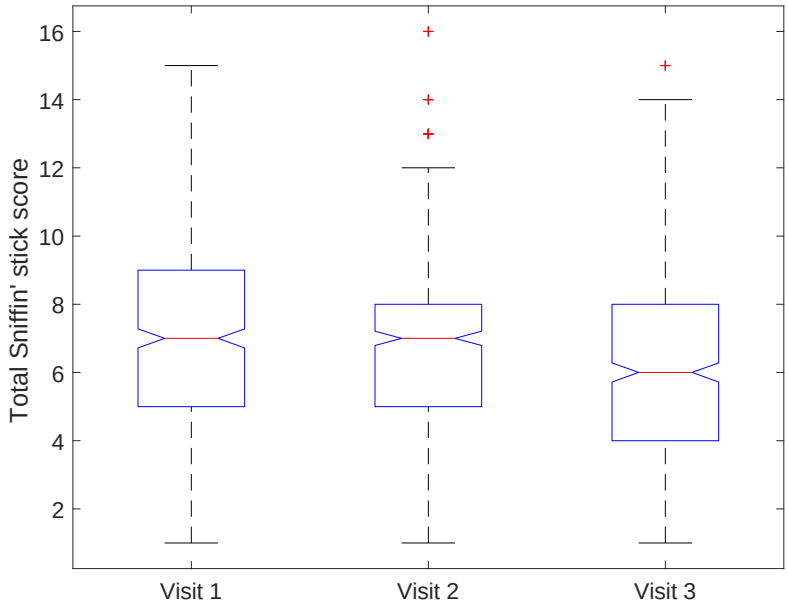
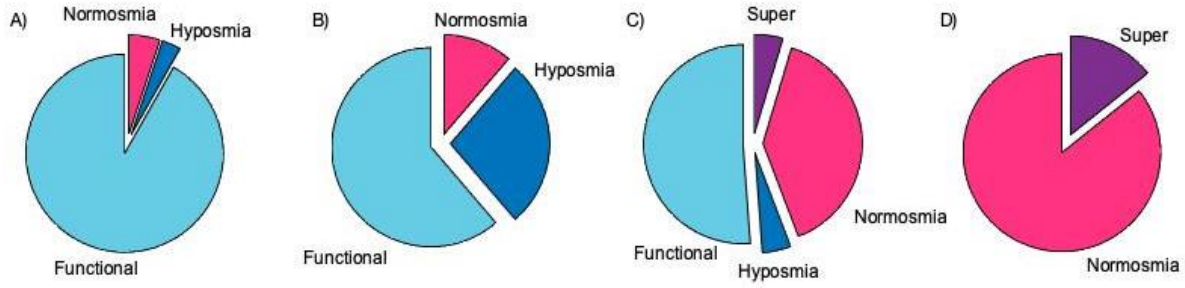


Figure 12 | Change in smell status between baseline and 36 month visits in individuals with PD.



Each pie chart relates to individuals with a given smell status at baseline (Pie chart A = baseline functional anosmia, Pie chart B = baseline hyposmia, Pie chart C = baseline normosmia, Pie chart D = baseline super smell) and their smell status at their 36 month visit i.e. of the individuals with hyposmia at baseline (Pie chart B), 28% continued to have hyposmia at their 36 month visit whereas 62% developed functional anosmia and 11% reverted to normosmia.

3.5.2 CLASSIFICATION: Distinguishing PD, RBD and healthy controls using sense of smell

3.5.2.1 *Applying different cut-off values to the 16-item total Sniffin' stick score, the AUC (95% confidence interval (CI)) for distinguishing RBD from controls was 0.84 (0.80-0.87); PD from controls, 0.90 (0.88-0.92) and PD from RBD 0.57 (0.53-0.61).*

The optimal operating point of the ROC curve allowed the identification of the optimal cut-off value for each pairwise comparison as illustrated in Table 8. The cut-off values were then applied to the classification validation datasets formed using PD (Tracking) and RBD (Marburg) data combined with longitudinal control (Discovery) data, which was treated as independent of baseline control data (Table 9).

Table 8 | The optimal cut-off values and their associated sensitivity, specificity, PPV and NPV for distinguishing PD and RBD from controls and each other.

Comparison	Optimal cut-off	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RBD v. Controls	≤9	71.4	87.9	81.9	79.9
PD v. Controls	≤10	86.9	78.3	91.9	67.7
RBD v. PD	≤14	99.7	2.9	79.1	70.0

Table 9 | Application of identified cut-off values to classification validation datasets.

Comparison	Cut-off applied	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RBD v. Controls	≤9	75.7	87.5	84.9	79.6
PD v. Controls	≤10	84.7	77.5	97.7	31.0

3.5.2.2 *Comparing machine learning approaches on entire (unbalanced) datasets (RBD/control, PD/control and RBD/PD), random forests provided close to the highest accuracy values.*

Using 10 fold cross validation and the raw answers provided for each of the 16 Sniffin' sticks, random forests distinguished RBD from controls, PD from controls and RBD from PD with 80.1%, 83.7% and 77.5% accuracy. Comparatively, boosted tree ensemble provided respective accuracies of 80.7%, 84.3% and 77.3% (Table 10).

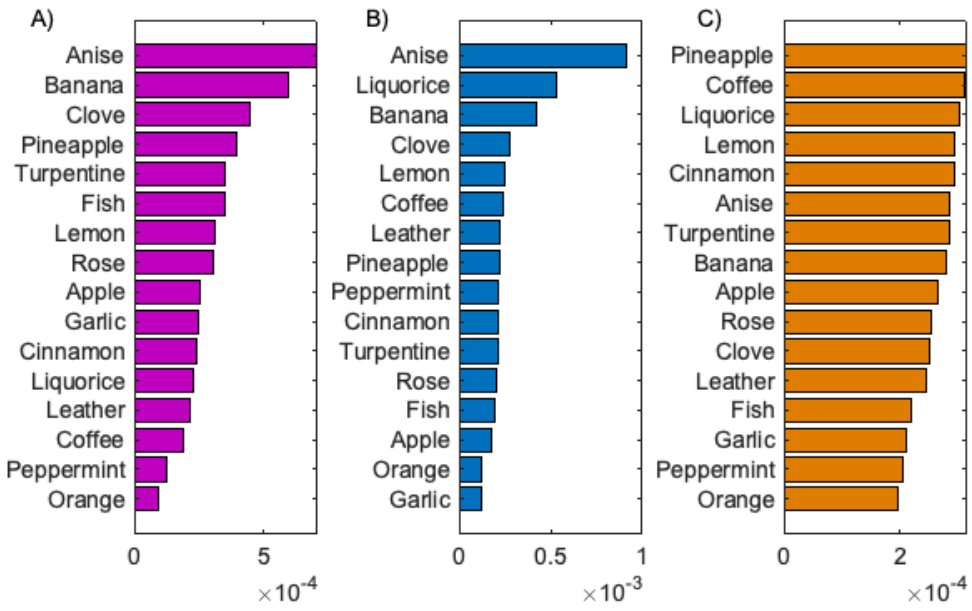
Table 10 | Accuracies associated with different machine learning approaches for distinguishing disease groups on the basis of smell.

Type	Specifics	RBD v. Controls	PD v. Controls	RBD v. PD
Logistic Regression	Logistic Regression	78.2	83.9	77.0
Support Vector Machines	Linear SVM	80.1	84.8	78.7
	Quadratic SVM	78.7	84.1	77.9
	Cubic SVM	76.4	82.1	70.2
k-Nearest Neighbours	Fine KNN	65.7	73.9	73.8
	Medium KNN	70.4	79.6	78.2
	Coarse KNN	69.9	80.8	78.7
	Weighted KNN	71.5	80.3	77
Ensemble	Boosted Trees	80.7	84.3	77.3
	Random Forests	**80.1**	**83.7**	**77.5**

3.5.2.3 *Using a random forests machine learning approach and all 16 Sniffin' sticks the overall AUC (95% CI) was 0.86 (0.85-0.87) for distinguishing RBD from controls, 0.89 (0.88-0.90) for distinguishing PD from controls and 0.62 (0.60-0.63) for distinguishing RBD from PD.*

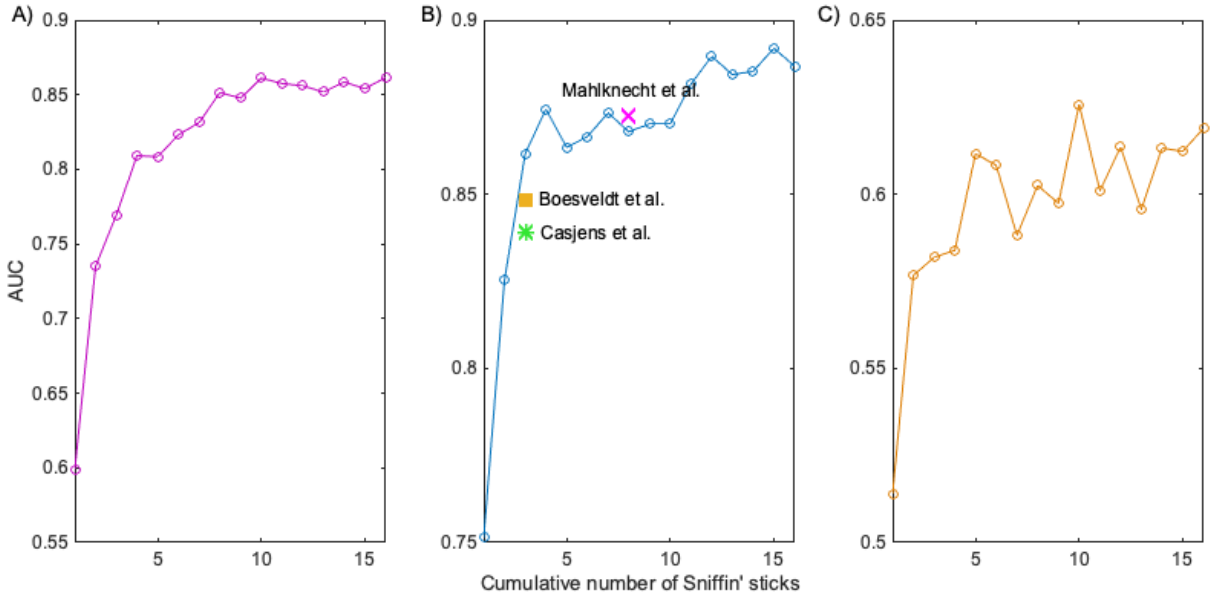
Applying a single randomly selected trained model to the classification validation datasets, similar AUCs were obtained; specifically, 0.88 for PD versus controls, 0.84 for RBD versus controls and 0.58 for RBD versus PD. The predictor importance and effect of sequential feature selection on model accuracy for each comparison are shown in Figure 13 and Figure 14.

Figure 13 | Comparative mean predictor importance of individual Sniffin' sticks in the classification of individuals according to groups.



Distinguishing A) RBD from controls, B) PD from controls, C) RBD from PD.

Figure 14 | Improvement in AUC with the sequential addition of Sniffin' sticks by order of predictor importance for distinguishing A) RBD from controls, B) PD from controls, C) RBD from PD.



Green asterix: Casjens et al. (3 stick = Peppermint / Coffee / Anise) combination; Orange square: Boesveldt et al. (3 stick = Cinnamon / Liquorice / Anise) combination; Magenta cross:

Mahlknecht et al. S8 (8 stick = Cinnamon / Peppermint / Banana / Liquorice / Coffee / Pineapple / Rose / Anise) combination.

3.5.2.4 *Using a machine learning approach and the selected abbreviated Sniffin' stick number for each comparison (the minimum number of sticks allowing for the maximum increment in AUC being selected), the overall AUC (95% CI) for distinguishing RBD from controls using 4 sticks was 0.81 (0.80-0.82), PD from controls using 3 sticks 0.86 (0.85-0.87) and RBD from PD using 5 sticks 0.61 (0.60-0.63).*

The respective Sniffin' stick combinations were Anise / Banana / Clove / Pineapple (RBD versus controls), Banana / Liquorice / Anise (PD versus controls) and Pineapple / Coffee / Liquorice / Lemon / Cinnamon (PD versus RBD). For each abbreviated Sniffin' stick combination, the improvement in AUC compared to using a single Sniffin' stick was >80% the maximal possible improvement with the addition of the remaining 15 sticks.

3.5.2.5 *Using a "brute force" approach (in the absence of cross validation), with an equivalent number of sticks for each comparison and the number correctly identified, the maximum AUC (95% CI) for distinguishing RBD from controls was 0.82 (0.78-0.85), PD from controls 0.87 (0.84-0.89) and PD from RBD 0.64 (0.60-0.67) (Figure 15).*

The respective Sniffin' stick combinations were Banana / Clove / Anise / Fish (RBD versus controls) out of 1820 possible 4 stick combinations, Banana / Liquorice / Anise (PD versus controls) out of 560 possible 3 stick combinations and Orange / Peppermint / Liquorice / Coffee / Anise (PD versus RBD) out of 4368 possible 5 stick combinations. The AUC distributions associated with all possible stick combinations for each

comparison are shown in Figure 15. The 3 Sniffin' stick combinations described by Boesveldt et al. and Casjens et al. yielded AUCs of 0.83 (0.81-0.86) and 0.84 (0.82-0.86) for distinguishing individuals with PD from controls respectively.[122, 130]

The machine learning approach was compared to the "brute force" approach of evaluating every possible Sniffin' stick combination for a given number of Sniffin' sticks, using the classification validation datasets as shown in Table 11.

Figure 15 | Histograms illustrating the distribution of AUC values associated with all possible 4, 3 and 5 Sniffin' stick combinations for the separation of individuals with A) RBD from controls, B) PD from controls and C) RBD from PD respectively.

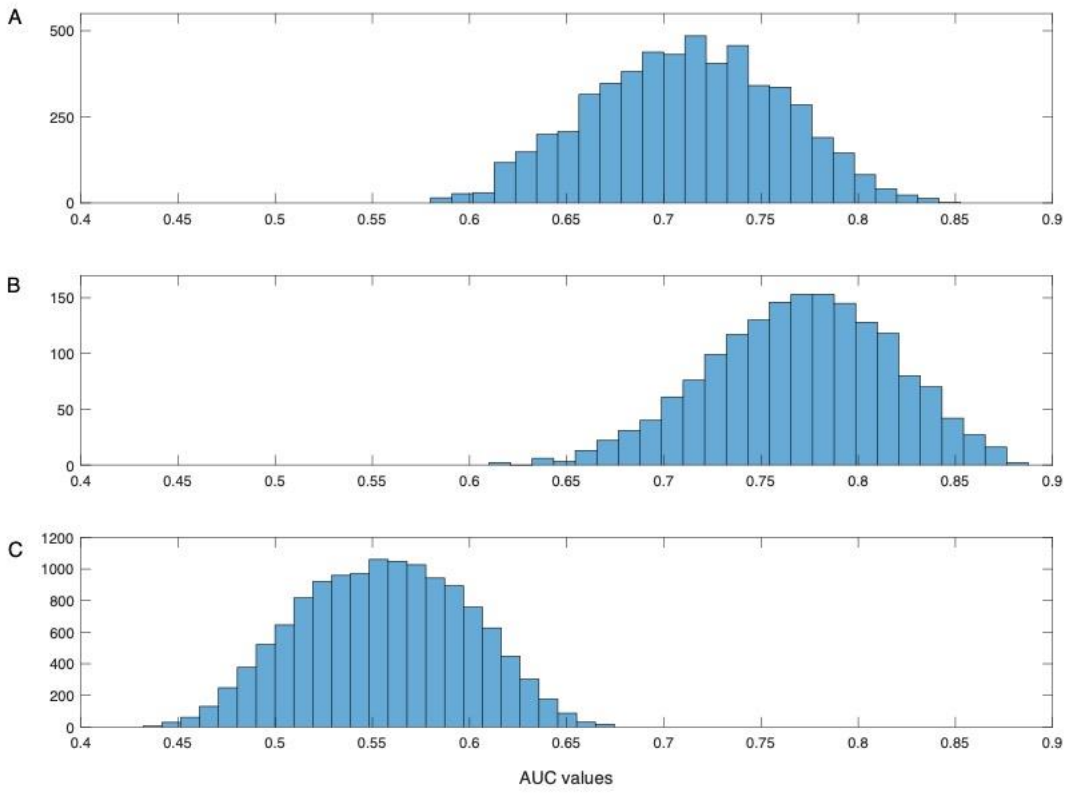


Table 11 | Comparison of machine learning and "brute force" approaches in distinguishing groups using the classification validation datasets.

Comparison	Machine learning approach	"Brute force" approach
RBD v. Controls	0.78 (0.66-0.88)	0.80 (0.69-0.88)
PD v. Controls	0.84 (0.78-0.88)[a]	0.85 (0.79-0.89)[a]
RBD v. PD	0.57 (0.47-0.66)	0.54 (0.45-0.63)

[a]The same optimal 3 Sniffin' stick combination was identified through both approaches.

3.5.3 STRATIFICATION: Delineating sense of smell in PD and RBD

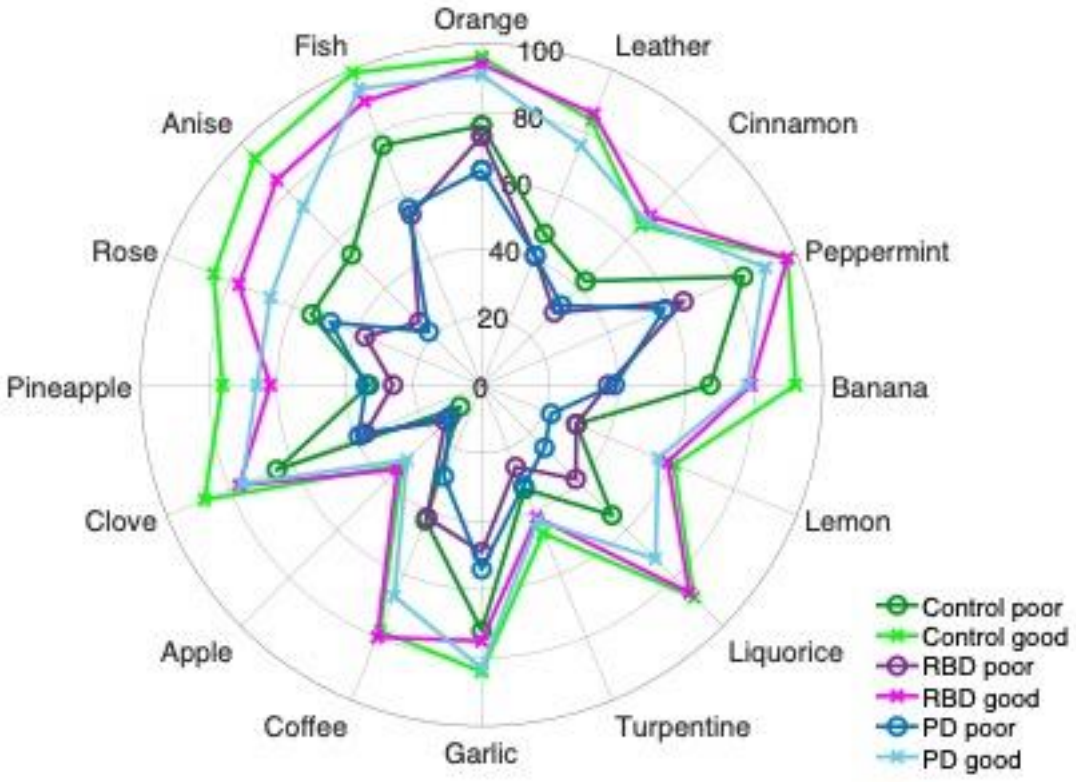
3.5.3.1 *Independent of disease group and smell status there were differences in the rates of correct identification between Sniffin' sticks.*

Orange, Peppermint and Fish had the highest rates of correct identification (poor smell: ≥55%, good smell: ≥90%, overall: ≥70%) compared to Lemon, Turpentine and Apple with the lowest (poor smell: ≤31%, good smell: ≤59%, overall: ≤36%).

3.5.3.2 *Overall, compared to individuals with PD, controls of the same smell status demonstrated higher rates of correct identification for each individual Sniffin' stick (Figure 16 & Table 12).*

The greatest differences in rates of correct identification between controls and individuals with PD of the same smell status were for Banana, Liquorice and Anise ($p<0.001$).

Figure 16 | A spider web plot illustrating the rates of correct identification for each Sniffin' stick by smell status and disease group.



Each line radiating out from the centre represents a different Sniffin' stick and the distance of each point from the centre indicates the proportion of individuals who correctly identified the respective stick. The maximal radius of the web denotes a 100% rate of correct identification. Independent of disease group or smell status, Apple was poorly identified. Anise, Banana and Liquorice were the three sticks with the greatest differences in rates of correct identification between controls and individuals with PD.

Table 12 | Sniffin' stick identification in individuals with PD and controls by smell status.

		Poor smell[a]			Good smell[b]		
		Percentage correctly identified			Percentage correctly identified		
Stick number	Stick smell	Controls	PD	p[c]	Controls	PD	p[c]
1	Orange	76	63	0.06	96	91	0.02
2	Leather	48	41	0.37	84	76	0.06
3	Cinnamon	43	33	0.15	66	68	0.59
4	Peppermint	83	58	<0.01	97	90	<0.001
5	Banana	67	39	<0.001	92	78	<0.0001
6	Lemon	30	22	0.20	61	56	0.23
7	Liquorice	54	26	<0.0001	88	72	<0.0001
8	Turpentine	33	32	0.92	47	43	0.41
9	Garlic	72	54	0.02	84	83	0.77
10	Coffee	43	29	0.04	78	67	0.01
11	Apple	9	13	0.38	33	31	0.63
12	Clove	65	39	<0.001	88	76	<0.01
13	Pineapple	33	34	0.89	76	66	0.01
14	Rose	54	48	0.42	85	67	<0.0001
15	Anise	54	22	<0.0001	94	74	<0.0001
16	Fish	76	56	<0.01	99	94	0.01

[a]Individuals with hyposmia or functional anosmia. [b]Individuals with normosmia or super smell. [c]p value determined using a chi squared test.

3.5.3.3 *Comparing machine learning approaches on the entire PD poor smell/control good smell (unbalanced) dataset, random forests provided close to the highest accuracy values.*

Using 10 fold cross validation and the raw answers provided for each of the 16 Sniffin' sticks, random forests distinguished individuals with PD and poor smell from controls with good smell with 95.5% accuracy, compared to the 96% accuracy associated with Cubic SVM (Table 13).

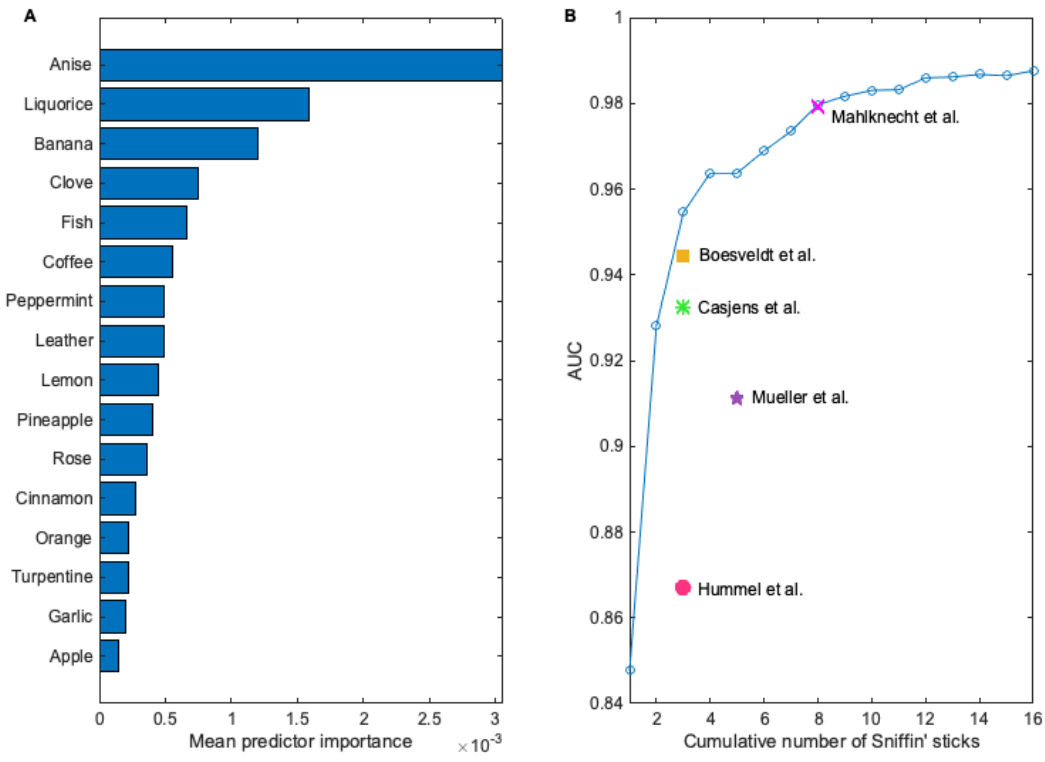
Table 13 | Accuracies associated with different machine learning approaches for the detection of PD poor smell.

Type	Specifics	Accuracy (%)
Logistic Regression	Logistic Regression	94.2
Support Vector Machines	Linear SVM	95.9
	Quadratic SVM	95.4
	Cubic SVM	96.0
k-Nearest neighbours	Fine KNN	90.7
	Medium KNN	90.5
	Coarse KNN	86.7
	Weighted KNN	90.7
Ensemble	Boosted Trees	94.3
	Random Forests	**95.5**

3.5.3.4 *Using a random forests machine learning approach, the overall AUC associated with models trained using all 16 sticks to distinguish poor from good smell was 0.99.*

The predictor importance of each Sniffin' stick was derived from each trained random forests model using Gini's diversity index criteria for binary splitting and then averaged across all 5340 models. Anise, liquorice and banana were identified as the top three Sniffin' sticks (Figure 17A).

Figure 17 | A) The comparative average predictor importance of each Sniffin' stick. B) The effect of adding Sniffin' sticks in order of predictor importance on AUC values compared with other abbreviated Sniffin' stick combinations.



Blue line with unfilled dots: AUC and cumulative Sniffin' stick number, added in order of predictor importance; Pink filled in dot: Q stick Hummel et al. (3 stick) combination; Green

asterix: Casjens et al. (3 stick) combination; Orange square: Boesveldt et al. (3 stick) combination; Purple pentagon: Mueller et al. BSIT (5 stick) combination; Magenta cross: Mahlknecht et al. S8 (8 stick) combination.

3.5.3.5 *The overall AUC derived from models trained using the top 3 sticks, as identified through predictor importance, was 0.95 in the development dataset, equating to a sensitivity, specificity, PPV and NPV of 90%, 92%, 92% and 90% respectively, assuming a probability threshold set at 0.5 and remained ≥0.90 when a single randomly selected trained model was applied to internal and external PD and RBD validation sets (Table 14).*

Within the development dataset, evaluation of 3 stick model accuracy according to the age and sex groups used to categorize Sniffin' stick evaluated sense of smell, revealed subgroup AUCs largely in keeping with the overall AUC of 0.95 (Table 15). The lowest AUCs were associated with the 31 to 40 age group, independent of sex where they remained ≥0.80. Sex differences were most pronounced within the 51 to 60 age group where the AUC was 0.87 in females compared to 0.96 in males.

Table 14 | The comparative performance of random forests models trained using the top 3 machine learning algorithm (MLA)-identified sticks and those previously described by Hummel et al. as part of the 3 Q sticks combination for the detection of poor smell.

	AUC (95% CI)		
	16 sticks	MLA-identified 3 sticks	3 Q sticks[126]
Development			
Overall PD poor smell model accuracy[a]	0.99 (0.99-0.99)	0.95 (0.95-0.96)	0.87 (0.86-0.88)
Validation			
Internal: RBD baseline[b]	-	0.90 (0.85-0.94)	0.83 (0.76-0.88)
External: PD Tracking[b]	-	0.90 (0.85-0.93)	0.81 (0.76-0.86)
External: RBD Marburg[b]	-	0.95 (0.82-0.99)	0.82 (0.58-0.97)

[a]Calculated across 5340 trained models using data excluded from the training process. [b]For each 3 stick combination (MLA-identified or Q sticks) one model (out of the 5340 models trained) was randomly selected and applied to the validation datasets. The CIs calculated for the development and validation datasets are not directly comparable; in the development dataset they are calculated using a bootstrapping approach whereas in the validation dataset each set of clinical data is associated with a single prediction across which, the CI is calculated.

Table 15 | Subgroup analysis demonstrating the effect of age and sex on MLA-trained 3 Sniffin' stick model accuracy.

Overall PD poor smell model accuracy		Sex		
		OVERALL	MALE	FEMALE
Age group[a]	OVERALL		0.96 (0.95-0.96)	0.95 (0.93-0.96)
	31 to 40	0.82 (0.67-0.92)	0.80 (0.56-0.95)	0.83 (0.58-0.96)
	41 to 50	0.95 (0.92-0.97)	0.95 (0.89-0.98)	0.95 (0.91-0.98)
	51 to 60	0.94 (0.92-0.95)	0.96 (0.95-0.97)	0.87 (0.83-0.91)
	61 to 70	0.98 (0.97-0.98)	0.98 (0.97-0.99)	0.98 (0.96-0.99)
	71 to 80	0.94 (0.93-0.95)	0.92 (0.90-0.94)	0.98 (0.96-0.98)
	81 and over	0.96 (0.92-0.98)	0.94 (0.89-0.97)	1.00 (1.00-1.00)

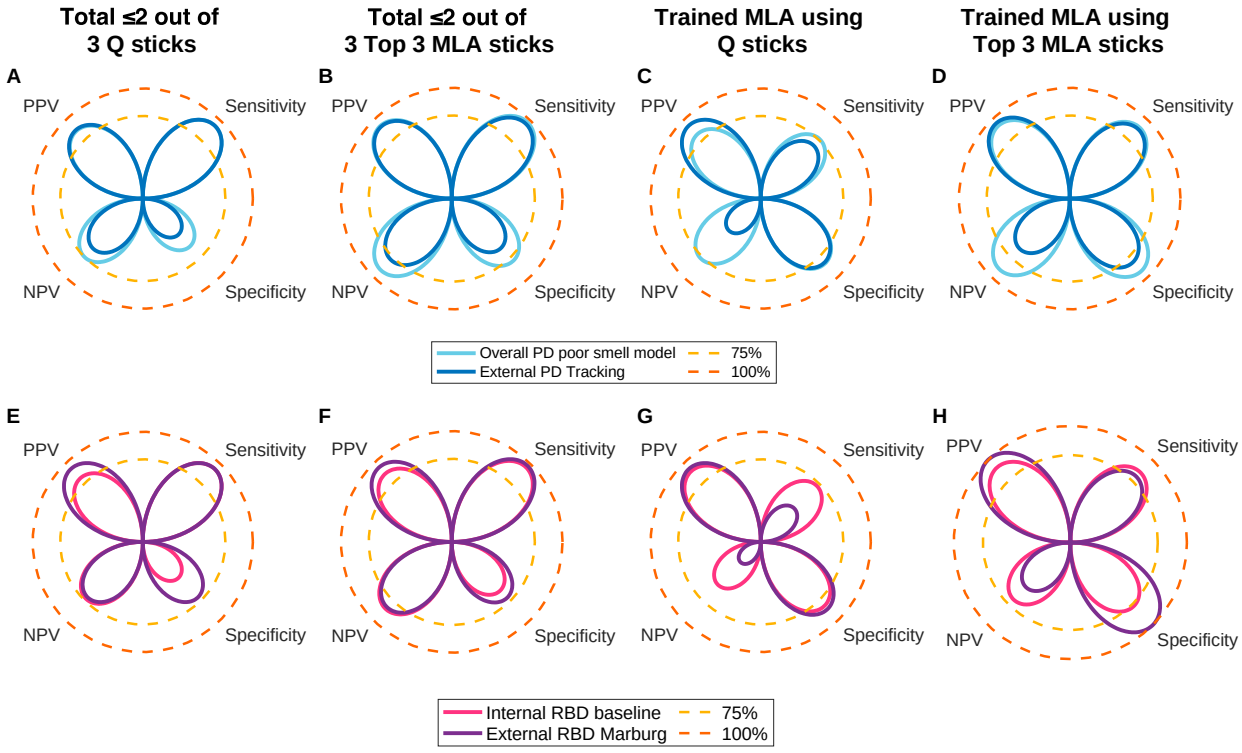
AUCs were calculated across 5340 trained models using data excluded from the training of each model. [a]All individuals within the 21 to 30 age group had a good sense of smell and so AUC value calculation was not possible.

3.5.3.6 *Random forests models trained using the 3 Q Sniffin' sticks combination (Coffee, Clove and Rose) described by Hummel et al. for the detection of poor smell, and those trained using the 3 stick combinations described by Boesveldt et al. (Cinnamon, Liquorice and Anise) and Casjens et al. (Peppermint, Coffee and Anise), originally for the separation of individuals with PD from controls, were outperformed by the machine learning algorithm (MLA)-identified 3 stick models in distinguishing poor from good smell (Table 14 and Figure 17b).*[126]

Similarly when using an absolute cut-off of ≤2 out of 3 correctly identified Sniffin' sticks the MLA-identified 3 stick combination outperformed the 3 Q sticks combination (Figure

18). When validated against the composite validation dataset (combining Discovery RBD, Marburg RBD and Tracking PD datasets), the AUC associated with our top 3 stick combination was 0.90 (0.89-0.91) compared to 0.68 (0.67-0.70) for the worst predicted 3 stick combination (Turpentine / Garlic / Apple). No other 3 stick combination was statistically better than our 3 stick (Anise / Liquorice / Banana) combination, in detecting poor smell ($p>0.05$). There was a significant difference ($p<0.001$) between the 3 stick combinations that included at least one of anise / liquorice / banana compared to those that did not include any of the 3 aforementioned sticks.

Figure 18 | Petal plots comparing machine learning and cut-off threshold approaches using the top 3 MLA-identified and Q sticks Sniffin' sticks combinations.



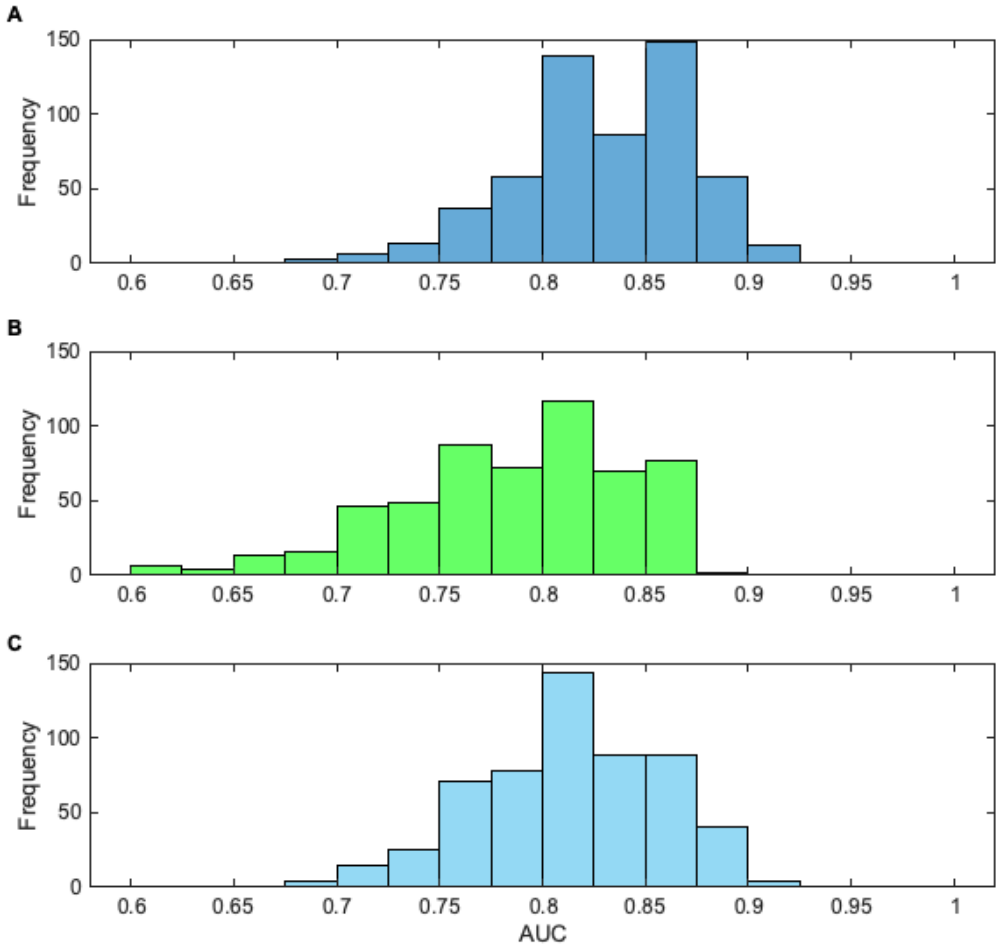
Accuracy values associated with the application of an absolute cut-off of ≤2 out of 3 sticks (A,B,E,F) or a trained random forests model (C,D,G,H) using the top 3 MLA-identified sticks (B,D,F,H) or empirically chosen Q sticks (A,C,E,G), described by Hummel et al.[126] to internal

(Discovery) and external PD (A,B,C,D) and RBD (E,F,G,H) datasets. The petal tip signifies the sensitivity, specificity, PPV and NPV for each approach and stick combination, with larger petal sizes indicating greater accuracies. The top 3 MLA-identified sticks outperform the equivalent number of Q sticks when using either an absolute cut-off (B,F versus A,E) or a trained random forests model (D,H versus C,G).

3.5.3.7 *For the detection of poor smell, models trained using PD poor smell/control good smell data were associated with greater accuracy than those trained using PD poor smell/PD good smell or Control poor smell/Control good smell data.*

The combination of PD poor smell/control good smell data led to an overall AUC value of 0.95, compared to 0.81 when the training data comprised control poor smell/control good smell data and 0.86 for PD poor smell/PD good smell data. The combination of PD poor smell/control good smell training data again proved superior when all 560 possible 3 stick combinations were evaluated for the detection of poor smell in individuals with RBD ($p<0.0001$) (Figure 19).

Figure 19 | The comparative effect of different training data on the detection of poor smell.



A) PD poor smell/control good smell, B) Control poor smell/control good smell and C) PD poor smell/PD good smell training data. The three distributions differ significantly ($p<0.0001$).

3.5.3.8 *The presence of an RBDSQ≥5 and poor smell as predicted using the abbreviated MLA-identified 3 Sniffin' stick test was associated with a sensitivity of 65%, specificity of 100%, PPV of 100% and NPV of 30% in distinguishing individuals with RBD from controls (Table 16).*

The RBDSQ≥5 and 3 Sniffin' stick detected poor smell, two-step screening test did not benefit from the addition of age≥55.

Table 16 | Comparative accuracy values for distinguishing individuals with RBD from controls using age, the RBD screening questionnaire and presence of poor smell (as identified through the MLA-identified 3 Sniffin' stick test) alone and in combination.

	Sensitivity	Specificity	PPV	NPV
Age[a]	88%	14%	87%	16%
RBDSQ[b]	97%	80%	97%	82%
Sniffin'[c]	67%	80%	96%	27%
Age[a] + Sniffin'[c]	63%	83%	96%	26%
Age[a] + RBDSQ[b]	86%	89%	98%	49%
RBDSQ[b] + Sniffin'[c]	65%	100%	100%	30%
Age[a] + RBDSQ[b] + Sniffin'[c]	62%	100%	100%	28%

[a]Age≥55 [b]RBDSQ≥5 [c]Poor smell as predicted by the MLA-identified 3 stick test

An increase in accuracy values is seen with the addition of the RBDSQ to the MLA-identified 3 Sniffin' stick test, with superior specificity and PPV compared to the RBDSQ alone.

3.6 Discussion

Summary of results:

1. Impaired olfaction is common both in PD and RBD
2. Smell status may change during longitudinal follow up
3. Individuals with RBD and PD can be distinguished from controls on the basis of a 16 Sniffin' stick test through the application of cut-off values or trained machine learning algorithms
4. Similar results are obtained using an abbreviated number of Sniffin' sticks to distinguish disease groups
5. Sniffin' stick identification differs between and within disease groups
6. An abbreviated 3 stick test (Anise / Liquorice / Banana) can be used to identify individuals with a poor sense of smell with excellent accuracies, confirmed on internal and external validation, and outperforming previously described 3 stick combinations for the detection of poor smell
7. High specificities are obtained through the application of a two-step screening test (RBDSQ≥5 and poor smell identified on 3 stick testing) in separating individuals with RBD from controls

3.6.1 Impaired olfaction is common both in PD and RBD

Including both development and validation cohorts, the results described relate to Sniffin' stick assessments from a total of 1933 individuals with isolated RBD, PD and controls; the largest study involving the assessment of smell in individuals with pre-manifest and manifest PD to date. Poor sense of smell on objective testing, including those categorised as hyposmic or functional anosmic, was present in 80% of individuals with Parkinson's, 72% with isolated RBD and 15% of controls. The higher rate of

impaired olfaction in individuals with isolated RBD in our study compared to others is likely to relate to differences in definition, dependent on the chosen smell assessment methodology, as well as the relatively recent introduction of Sniffin' stick age- and sex-specific normative data for individuals above the age of 55.[78, 115, 132, 133] Comparatively, problems with sense of smell were self-reported in 61% of individuals with PD, 58% of those with RBD and 11% of controls (Table 7). A loss of awareness of hyposmia despite greater deficits on objective smell testing, has previously been described in individuals with PD and mild cognitive impairment, compared to those with normal cognition.[134] In our study, applying a MoCA<26 as the cut-off definition for mild cognitive impairment, in keeping with MDS Task Force Level I criteria, there was no significant difference (p=0.26) in subjectively reported poor smell or objectively assessed poor smell (p=0.31) between mild cognitive impairment and normal cognition PD groups.[135]

3.6.2 Smell status may change during longitudinal follow up

93% of individuals with RBD and 94% of those with PD, and poor smell at baseline continued to have poor smell after 3 years of follow up (Figure 10 and Figure 12). The observed stability of poor smell identified at baseline is in keeping with other longitudinal studies of olfaction in RBD and PD.[132, 136, 137] In contrast, a baseline good sense of smell was maintained in 62% of those with RBD and 48% of those with PD. Though representing a small proportion of the overall disease groups, the prognostic bearing of good smell stability may benefit from further evaluation.

3.6.3 Individuals with RBD and PD can be distinguished from controls on the basis of smell

Comparable AUC values (0.84, 0.90, 0.57 versus 0.86, 0.89, 0.62) for distinguishing RBD from controls, PD from controls and RBD from PD respectively were obtained through the application of cut off-values to the total Sniffin' stick score and by using a machine learning approach, utilising the individual answers to each of the 16 Sniffin' sticks. The machine learning approach provided insight into the relative importance of each stick for a given classification prediction (Figure 13). Anise was the stick with the greatest importance in distinguishing individuals with RBD and PD from controls and featured in abbreviated Sniffin' stick combinations proposed by Boseveldt et al., Casjens et al., and Mahlknecht et al.[122, 123, 130] Consistent with the poor AUCs obtained when attempting to distinguish individuals with RBD from PD using Sniffin' stick assessments, there was little difference in predictor importance between sticks; the predictor importance of the lowest ranked stick (Orange) was 60% that of the highest ranked stick (Pineapple). The difficulty in separating individuals with RBD from PD on the basis of smell is not unexpected given their temporal association and supports the recapitulation of a PD-specific smell deficit, in individuals with RBD.

In distinguishing individuals with RBD from controls, our AUC of 0.84 (95% CI 0.80-0.87), derived by applying a cut-off threshold to the total Sniffin' stick odour identification score, was considerably lower than the AUC of 0.92 reported by Campabadal et al. using the same approach.[121] In their smaller cohort, individuals with RBD (n=21) were older (mean age 71.8 compared to 64.4) and of longer disease duration (mean (standard deviation(SD)): 4.6 (3.3) compared to 1.4 (1.8)); the higher rate of hyposmia

in their RBD group (91% versus 67%) likely accounting for their greater ease of separation from controls.

In contrast, separating individuals with PD from controls, the AUC (0.90) and optimal cut-off point (≤10) in our study closely matched those reported by Boesveldt et al. (AUC 0.91, cut-off 10.5).[130]

3.6.4 Abbreviated smell tests yield similar classification accuracies

The minimum number of sticks necessary to provide a classification AUC comparable with that associated with all 16 Sniffin' sticks were chosen for each pairwise comparison. 4, 3 and 5 stick AUCs of 0.81, 0.86 and 0.61 in distinguishing RBD from controls, PD from controls and RBD from PD using a machine learning approach (Figure 14) were similar to those associated with all 16 sticks and with the maximal respective AUCs of 0.82, 0.87 and 0.64, obtained for an equivalent stick number using a "brute force" cut-off threshold approach, where the AUC for every single possible combination of a given number of sticks was calculated (Figure 15). Comparable results for each approach were obtained when assessed in the classification validation datasets (Table 11). At a group level, whilst there is computational benefit in using a machine learning approach, particularly for the identification of the number of sticks to comprise the abbreviated smell test, once the optimal abbreviated stick combination has been identified, similar classification results can be obtained through the simple application of a cut-off threshold to the correct number of sticks identified. The added potential benefit of using a trained machine learning algorithm to classify individuals is the derivation of a probability of a given class e.g. the probability of PD versus controls, specific to each individual; its potential role in disease stratification warrants further investigation.

3.6.5 Sniffin' stick identification differs between and within disease groups

In keeping with previous reports, Orange, Peppermint and Fish[138] were associated with the highest rates of correct identification and Lemon, Turpentine and Apple,[123] the lowest (Figure 16). In all but the PD and RBD super smeller group (where it was the second worst), Apple was the worst identified stick; its rate of 27% correct identification in normosmic controls far below the >75% target for odourant selection applied at the inception of the Sniffin' sticks test, albeit created using data from much younger controls.[80] The three sticks with the greatest differences in rates of correct identification between controls and individuals with PD, with poor and good smell were Banana, Liquorice and Anise; the same three sticks were identified as having the top predictor importance for distinguishing PD from controls and for identifying PD specific poor smell (Table 12).

3.6.6 An abbreviated 3 stick test (Anise / Liquorice / Banana) detects PD specific poor smell with excellent accuracies

Aiming to derive an abbreviated Sniffin' stick test for the detection of PD specific deficits in smell, assuming a continuum wherein isolated RBD, prodromal to PD, lies between controls and PD, positioned at extremes of the spectrum, models were trained with data from controls with good smell (normosmia and super smellers) and individuals with PD and poor smell (hyposmia and functional anosmia). The combination of PD poor smell/control good smell training data proved superior to control good smell/control poor smell data and PD good smell/PD poor smell data at identifying poor smell in individuals with RBD, pointing towards a pattern of smell loss specific to PD that is recapitulated in RBD; itself, prodromal to PD (Figure 19).

The similarity in olfactory loss in RBD, compared to PD, bears close homology to that of other non-motor symptoms; namely cognition, depression, anxiety, apathy, impulse control behaviours, sleep disturbance and autonomic dysfunction.[34, 63, 64, 100, 139-141] Along with similarities in their motor profile, albeit to a lesser degree, the established picture is one wherein prodromal parkinsonism, within which RBD lies, transitions to, yet is contiguous with, motorically manifest PD.[18, 63, 64]

Although our models were not adjusted for age or sex, both variables were used to determine smell class by referencing age- and sex-specific normative data. The reliance of our models, purely on the raw answers provided for each stick by each participant, to classify individuals as having poor or good smell, may explain the 16 stick AUC of 0.99 (as opposed to 1.00), derived using test data, mutually exclusive from that used to train the models. Subgroup evaluation of model performance by age and sex revealed largely equivalent AUCs; lower AUCs were observed in the 31 to 40 age group and likely relate to the limited size of the subgroup dataset (Table 15).

A relative reduction in AUC improvement was observed with the addition of a fourth stick to models utilising the top 3 sticks (Anise, Liquorice and Banana). Indeed 75% of the possible improvement in AUC obtained from using a single Sniffin' stick compared to all 16 Sniffin' sticks was realised through the addition of a further 2 sticks (Figure 17). Encouragingly, the AUC associated with our single randomly chosen Discovery PD poor smell/control good smell trained model remained >0.90 when evaluated in internal and external RBD and PD datasets (Table 14).

Previously described 3 and 5 Sniffin' stick combinations were outperformed by models incorporating Sniffin' sticks cumulatively in order of MLA-derived predictor importance.[122, 125, 126, 130] Compared to our top 3 stick model, 3 stick combinations described by Casjens et al. and Boesveldt et al. bore slightly lower AUCs; the similarity in AUCs perhaps in part explained by the commonality of Anise (the top ranked stick in order of predictor importance) to all three combinations.[122, 130] The Boesveldt et al. combination differed only from ours by the inclusion of Cinnamon as opposed to Banana.[130] Similarly, the Mahlknecht et al. 8 stick combination shared 5 out of 8 sticks with our top 8 stick combination and yielded an almost identical AUC.[123] Our top 3 MLA-identified stick model was additionally compared to the 3 Q sticks combination described by Hummel et al. specifically for the stratification of individuals by sense of smell.[126] In a head to head comparison of 1) our top 3 MLA-identified sticks to 2) the 3 Q sticks, and 3) trained random forest models using either 3 stick combination to 4) an absolute cut-off of ≤2 correctly identified sticks, the combination of our top 3 MLA-identified sticks and trained random forest models afforded the optimal specificity and PPV, albeit with slightly lower sensitivity and specificity values compared to those derived through the application of absolute cut-offs (Figure 18). Indeed, having trained random forests models using all possible 3 stick combinations, no other 3 stick combination was statistically better than our 3 stick (Anise / Liquorice / Banana) combination, in detecting poor smell ($p>0.05$).

3.6.7 Poor sense of smell and an RBDSQ≥5 separates RBD from controls with high specificities

The RBD screening value of the top 3 MLA-identified stick test of poor smell, was assessed both alone and in combination with age≥55 and RBDSQ≥5.[31] Alone and

congruent with established literature, the RBDSQ demonstrated excellent sensitivity.[82] However, 100% specificity was obtained through the identification of individuals with both good smell on 3 stick testing and an RBDSQ<5, suggesting a promising role in community screening for subclinical RBD. No additional benefit was derived from the use of age≥55 as a screening criterion.

3.6.8 Limitations

3.6.8.1 *The need for an independent control cohort*

One of the main limitations in determining the generalisability of our models, was the lack of an independent control cohort. In its absence, independence of longitudinal control data from baseline control data was assumed during the creation of the somewhat artificial dataset with which our two-pronged screening test was evaluated.

3.6.8.2 *The potential limit to the generalisability of results from research participants with RBD to those with RBD as yet undiagnosed in the community*

The phenotype of individuals having acted on their own initiative to seek medical evaluation of their sleep disturbance, and eventual diagnosis of RBD, may be expected to be more severe than that of those who may otherwise remain undiagnosed in the community and be more challenging to detect. Having received a diagnosis of RBD and decided to partake in research, such individuals may be more likely to have higher levels of health literacy and with foreknowledge of their diagnosis, to answer affirmatively to the RBDSQ.

3.6.8.3 *The need to determine the cross-cultural application of results*

Future work will evaluate the use of the 3 Sniffin' stick/RBDSQ test in population screening for individuals with RBD, mindful that the community prevalence of RBD may impact upon the PPV. Despite the sizeable normative data reported by Oleszkiewicz et al., the population from which it was acquired was largely German based and so regional variations may be anticipated.[78] Mindful of cultural and regional variations in exposure to Anise and Licorice, both similar in smell, future application of our work, outside of Europe, to international datasets will allow the generalisability of our results, beyond the relatively homogeneous European cohorts described, to be assessed.

3.6.9 Concluding remarks

This chapter has explored the evaluation of sense of smell using Sniffin' stick assessments from the largest dataset involving individuals with prodromal and motorically manifest Parkinson's to date. Abbreviated Sniffin' stick tests have been derived for the purpose of classifying individuals according to disease group and categorising them according to presence of olfactory dysfunction, with results validated internally and externally in independent datasets. Having demonstrated high levels of accuracy in the detection of olfactory dysfunction with a 3 stick test, that has the feasibility to work within the time constraints of a busy clinical environment, a similar approach to the MDS-UPDRS III will be explored in the next chapter.

CHAPTER 4. Measuring motor function using existing gold standard clinical rating scales

4.1 Contextualisation

Having utilised machine learning to ascertain the relative predictor importance of each Sniffin' stick in the assessment of smell, this chapter applies a similar approach to the MDS-UPDRS parts I-III in order to delineate the relative importance of each item in the prediction of the total MDS-UPDRS III score and PD severity categorisation as well as in the determination of tremor dominant / postural instability and gait difficulty / indeterminate subtype. With a data driven approach, an abbreviated MDS-UPDRS III score is created, offering potential utility and insight into clinical practice.

4.2 Introduction

4.2.1 The UPDRS

The original UPDRS was first described by Fahn et al. in 1987. The total score, out of a maximum of 176, was derived through the summation of three subsections: the first section (maximum=16) comprised four questions that related to mentation, behaviour and mood; the second (maximum=52) related to activities of daily living and; the third (maximum=108) a motor examination.[142] A 2003 review of its utilisation amongst movement disorder specialists acknowledged its popularity as a gold standard reference scale yet recognised weaknesses in 1) the uncertainty of its relation to clinically pertinent disease severity categories 2) the absence of a minimal clinically relevant incremental difference that would discriminate between two treatments 3) the relative under-sampling of non-motor symptoms 4) the potential for cultural bias, particularly in Part II where, for example, dressing difficulty with buttons may be irrelevant in cultures

in which they are not used 5) the need for greater clarity in the instructions that guide administration of the scale.[143]

4.2.2 The MDS-UPDRS

Under the auspices of the MDS, the ensuing revision of the UPDRS, the MDS-UPDRS was released in 2008 and retained a similar structure to the original version, with parts I (maximum=52) and II (maximum=52) focussing on non-motor and motor experiences of daily living respectively and part III (maximum=132) encompassing the motor examination.[81] Each item is rated on a five point scale; normal (0 points), slight (1 point), mild (2 points), moderate (3 points) or severe (4 points). At its inception, the results of initial clinimetric testing, alongside the purported clinical relevance of each of its subparts, led the MDS task force to advocate reporting of subpart scores as opposed to a single total MDS-UPDRS score.[81] Detailed instructions, aimed at promoting standardised assessments, were incorporated into the scale. A greater emphasis was given to non-motor symptoms of PD. Formulas converting UPDRS II and III scores to the newer MDS-UPDRS were derived.[131] A 3.25 point improvement or 4.63 point worsening of the MDS-UPDRS III was found to be the minimum change associated with a clinically meaningful difference to patients.[144] Cut-off points, dividing individuals with PD into those with mild (Part I: ≤10; Part 2: ≤12; Part 3: ≤32), moderate (Part 1: ≥11 and ≤21; Part 2: ≥13 and ≤29; Part 3: ≥33 and ≤58) or severe (Part 1: ≥22; Part 2: ≥30; Part 3: ≥59) disease on the basis of their MDS-UPDRS subpart scores, were established using agreement between the Patient Global Impression of Severity and at least two of: Hoehn and Yahr staging, Clinical Impression of Severity Index for Parkinson's disease and Clinical Global Impression of Severity, as the "criterion of severity".[145]

4.2.3 Motor progression in PD

Analysis of 362 individuals with PD recruited to the Parkinson's Progression Markers Initiative (PPMI) Cohort revealed a yearly MDS-UPDRS Part II (self-reported motor severity) and III (researcher assessed motor severity) increment of 1.0 points and 2.4 points per year, respectively.[146] In addition to requiring participants to have either resting tremor or bradykinesia plus at least one other out of: resting tremor, bradykinesia and rigidity; or either asymmetric resting tremor or asymmetric bradykinesia, PPMI inclusion criteria included drug-naivety at the point of recruitment, a disease duration of less than 2 years from diagnosis and evidence of dopaminergic deficit on ^{123}I-ioflupane single-proton emission computed tomography imaging.[147] In comparison, participants recruited to the Discovery study were older, of longer disease duration, and had higher total and MDS-UPDRS I-III subpart scores at baseline.[34, 146] Despite the differences between the two cohorts, the yearly average increase in MDS-UPDRS Part II and III scores in the Discovery cohort were similar to those in PPMI at 1.2 and 2.2 points respectively.[127] The Tracking study, another UK PD cohort study, with close homology to the Discovery study (both studies involving community ascertained cases), also had broadly similar figures, with the MDS-UPDRS part II increasing by a mean of 1.4 points and the MDS-UPDRS part III increasing by a mean of 2.4 points per year.[127]

Of the two MDS-UPDRS parts that assess motor severity, the researcher assessed part III motor examination is more frequently used in clinical trials.[148] Whilst the MDS-UPDRS III subpart total can be considered to span all strata of disease severities, psychometric issues with the original UPDRS II and newer MDS-UPDRS II have been reported.[73, 149] Indeed, a significant floor effect in all but one item of the MDS-UPDRS II

was observed in individuals with early PD in the PPMI cohort and the degree of separation achieved was deemed modest.[73]

4.2.4 Inter- and intra-rater variability

Of Parts I-III, it is the motor examination score that is associated with the highest test-retest (intra-rater) reliability. In a study involving 400 patients recruited at different centres as part of a randomised, placebo-controlled trial of rasagiline for the symptomatic treatment of early PD, repeat screening and baseline pre-treatment assessments were performed by the same neurologist after a mean interval of 14.6 (SD 7.6) days. The intraclass correlation coefficient (ICC) for clinician-assessed part III was 0.90, compared to 0.85 for the self-rated part II.[150] In a small study involving 24 patients rated by three experienced movement disorder neurologists, within part III, individual items differed with respect to the degree of agreement between raters, with good-excellent agreement in the assessment of repetitive movements (ICC 0.70-0.92) and rest tremor (ICC 0.84) compared to poor agreement in the assessment of facial expression and speech (ICC 0.07 and 0.29 respectively).[151]

Whilst the ICC provides a measure of the between-patient variance as a proportion of the total variance, systematic differences between raters are not immediately evident. Indeed, inter-rater variability is particularly marked when comparing individuals with differing levels of experience as demonstrated in a study by Post et al. 50 participants were videoed undergoing their part III motor examination and subsequently rated by two nurse practitioners, two residents in Neurology, one movement disorders specialist and one senior movement disorders specialist; the latter, against whom all other raters were compared. The mean difference between raters, rating the same patients, varied from

1.7 to 5.4 points. There was a systematic difference in the ratings provided by the senior movement disorders specialist, which tended to be lower than those of the other raters; the authors postulated that their predominant exposure to patients with PD on the severe end of the disease spectrum may have been the cause. Worryingly, limits of agreement between raters were as large as 16 points, over 4 times the minimally clinically important difference that neuroprotective treatment trials are often powered to detect (Table 17).[72]

Significant within-subject measurement error has been noted when using the MDS-UPDRS III to track changes in PD progression over time. Data from 423 individuals with PD followed up for a median of 54 months as part of the PPMI study were used within a linear Gaussian state space model to determine the expected variance associated with true disease progression and the variance associated with short term fluctuations in disease and measurement error. Estimates of within-subject reliability for the MDS-UPDRS III total score were greater in the PD medication OFF, compared to ON state (0.50 versus 0.23). Factor evaluation revealed higher estimates of within-subject reliability for MDS-UPDRS III items relating to gait and posture (0.62) and rest tremor (0.43) compared to other forms of tremor, rigidity and bradykinesia (0.13-0.27).[152]

Inter-rater MDS-UPDRS III scoring disagreement does not only affect individuals with PD. In the PREDICT-PD study, which recruited healthy individuals over the age of 60 without known neurological disease, videoed MDS-UPDRS III assessments differed in their scoring by ≥5 points in 12% of participants rated independently by two raters.[153] Yet, despite the known weaknesses of the MDS-UPDRS III, in the absence of an alternative widely accepted measure of PD motor severity, its popularity persists.

4.2.5 The MDS-UPDRS in drug trials

To date, despite multiple neuroprotective treatment trials, none have been definitively proven to alter the progression of Parkinson's.[29, 154] Their failure, at least in part, has been blamed on the use of insensitive end points and outcome measures,[155, 156] of which the original total UPDRS, superseded by the MDS-UPDRS, or parts thereof, has been the most popular (Table 17). The entrenchment of the UPDRS as an outcome measure is fuelled by its acceptance by the Food and Drug Administration[157] as well as the European Medicines Agency[158] where it is promoted as a way of assessing the slowing of motor symptom progression in early PD, yet initiatives increasingly support the development of novel digital endpoints.[156]

Table 17 | Studies using the UPDRS / MDS-UPDRS or parts thereof as a primary outcome measure in clinical trials

Year	Trial	Primary outcome measure	Result
2020	STEADY-PD III: Multicentre double-blind, placebo-controlled, parallel group study of isradipine[159]	1. 4 point difference in mean total UPDRS score change from baseline to 36 months between isradipine and placebo groups	Failed: -0.27 point difference in total UPDRS score change between the two groups
2019	GDNF study: Single centre,	1. 20% change from baseline in practically	Failed: Decrease of 17.3 ± 17.6% (6.2 ± 7.1

	double-blind, placebo-controlled study of glial cell line-derived neuroptrophic factor (GDNF)[160]	defined OFF state MDS-UPDRS III	MDS-UPDRS III points) in GDNF group compared to 11.8 ± 15.8% (3.4 ± 4.3 points) in the placebo group. No statistically significant mean treatment difference
2019	LEAP study: Multicentre double-blind, placebo-controlled, delayed start study of levodopa with carbidopa[161]	1. 4 point difference in mean total UPDRS change from baseline to week 80 between early and delayed start groups	Failed: 1.0 point difference in total UPDRS score change between the two groups
2017	Exenatide study: Single centre, double-blind, placebo-controlled study of	1. Difference in change in MDS-UPDRS III scores in the practically defined OFF state at 60 weeks between the two groups	Promising: 3.5 point adjusted difference in MDS-UPDRS III between the two groups. Phase III trial ongoing to determine whether the effects observed relate

	exenatide[155]		to disease modification or a symptomatic effect
2015	NET-PD: Multicentre double-blind, placebo-controlled trial of pioglitazone in addition to rasagiline or selegiline[162]	1. 3 point difference in mean total UPDRS of each pioglitazone dose group compared to the placebo group	Failed: -1.83 and -1.12 difference in total UPDRS in 15mg and 45mg dose groups compared to placebo
2014	QE3 study: Multicentre double-blind, placebo-controlled study of coenzyme Q10[163]	1. Adjusted mean change in total UPDRS from baseline to final visit of 3 points between each coenzyme Q10 dose group compared to placebo	Failed: Adverse trend in active treatment group compared to placebo with total UPDRS change of 6.9 points in placebo group compared to 7.5 in 1200mg/day and 8.0 points in 2400mg/day coenzyme Q10 groups
2010	MitoQ study: Multicentre double-blind,	1. Change in total UPDRS score between the treatment groups over a	Failed: mean change of 4.94 points in the placebo group

	placebo-controlled trial of MitoQ[164]	12 month period	compared to 8.32 in the 40mg and 7.88 in the 80mg treatment arms
2009	ADAGIO study: Multicentre double-blind, placebo-controlled, delayed-start trial of rasageline[165]	1. Superiority to placebo in rate of change in total UPDRS between weeks 12 and 36 2. Superiority to delayed start in change in total UPDRS score between baseline and week 72 3. Non-inferiority to delayed start treatment in rate of change in total UPDRS score between weeks 48 & 72	Failed: All endpoints met with 1mg but not 2mg of rasagiline (where endpoint 2. was not met)

4.2.6 The MDS-UPDRS in clinical practice

Despite having had a role in supporting the use of many of the current symptomatic PD therapies available to date, the MDS-UPDRS is largely restricted to use within the field of research. Particularly in the NHS, time-limited appointment slots typically preclude the administration of the MDS-UPDRS III motor examination in its entirety, if at all. The non-standardised way in which motor examinations are performed and documented can make assessing individual treatment response challenging, particularly when a) the

interval between assessments can be long and b) different clinicians may be involved in seeing the same patient at different points in their care.

With the aim of deriving a brief PD rating scale that would be amenable for use in clinical practice, Hauser et al. described an abbreviated score, named the UPDRS-8. The “motor score” component of the UPDRS-8 comprised the sum of five individual UPDRS III item scores (gait and right/left assessments of upper limb rest tremor and finger tapping); each item having been chosen by researcher consensus.[166] There was a significant correlation between the “motor score” of the UPDRS-8 and the UPDRS III total (rho=0.77). Exploring the potential for remote administration of the UPDRS III, Abdolahi et al. used data from the Comparison of the Agonist pramipexole vs. Levodopa on Motor complications in Parkinson’s Disease (CALM-PD) study, to assess the effect of removing hands-on assessments of rigidity and postural instability, from the UPDRS III. The correlation of the resultant score (the sum of the remaining 21 individually rated items), with the UPDRS III total was 0.96 at baseline and remained above 0.90 longitudinally.[167] Since the release of the MDS-UPDRS, an MDS-UPDRS derived tremor-specific scale[168] has been proposed but to our knowledge, no other MDS-UPDRS III abbreviated score of overall PD motor function has been reported, nor data driven approach used, in the selection of constituent items.

4.2.7 PD subtyping

First conceptualised by the DATATOP (Deprenyl And Tocopherol Antioxidative Therapy Of Parkinsonism) study group, individuals with PD were divided into those with a predominance of postural instability and gait difficulty (PIGD) and those with tremor dominance, according to the ratio of their PIGD and tremor scores, each calculated from

the relevant individual UPDRS item scores.[169] Whilst subtypes are subject to a degree of temporal variability,[170-172] the PIGD phenotype has been associated with more rapid disease progression,[169, 173, 174] shortened survival,[175, 176] reduced levodopa responsiveness,[175] greater autonomic symptom severity,[177] more frequent falls,[178, 179] freezing,[180] cognitive decline,[181-185] and greater depression,[181] compared to the tremor dominant phenotype. Subtyping according to clinician-selected MDS-UPDRS item defined PIGD or tremor dominance remains one of the most popular methods of subclassifying individuals with PD; likely due to its relative ease of application compared to more sophisticated models, which, whilst providing greater insight are complicated by their multivariate requirement.[42, 127, 186]

4.3 Aims

The work in this chapter on the MDS-UPDRS III centres around three main aims:

4.3.1 PREDICTING the MDS-UPDRS III motor examination score

Items within parts I and II of the MDS-UPDRS (both of which may be administered remotely by a researcher or self-administered by a participant)[187] and items within the MDS-UPDRS III itself, are used to predict the MDS-UPDRS III (which involves time consuming in-person assessments by trained researchers).

4.3.2 PREDICTING PD disease severity and Tremor dominant / PIGD / indeterminate subtype

Items within parts I and II of the MDS-UPDRS are similarly used to train models aimed at classifying individuals according to MDS-UPDRS III motor severity (mild/moderate/severe) and subtype.

4.3.3 CREATING an abbreviated MDS-UPDRS III score

The average predictor importance of each item in MDS-UPDRS III, is used to inform the creation of an abbreviated MDS-UPDRS III score.

4.4 Methods

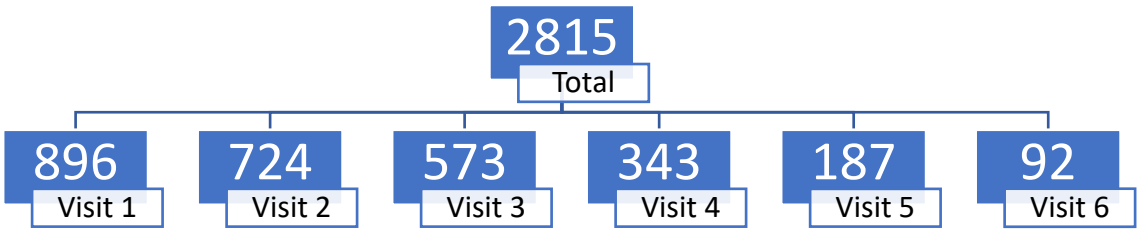
4.4.1 Clinical assessments

MDS-UPDRS parts I (self-reported non-motor symptom severity), II (self-reported motor symptom severity) and III (researcher assessed motor severity) were assessed at each clinic visit as detailed in the General Methods chapter. Unless otherwise specified, where MDS-UPDRS III assessments are bilateral, left and right sided items relating to the same task are considered independently.

4.4.2 Data analysed

At the point of data download (10/6/19), of the 929 participants with PD (Figure 1) with a probability of PD≥90% at their most recent follow up and with a disease duration of <3.5 years at the point of recruitment, 923 had contributed one or more complete (MDS-UPDRS parts I, II and III) sets of data at one or more visit, resulting in 2815 complete sets of data (Figure 20). Outside of longitudinal analyses, the 2815 sets of data were treated as independent of each other and used to train the prediction models.

Figure 20 | Complete sets of MDS-UPDRS parts I-III data contributed by participants with PD



4.4.3 Statistical analyses

4.4.3.1 *Descriptive statistics*

Simple descriptive statistics were used to compare changes in the items within parts I-III of the MDS-UPDRS over the first 3 visits (a 3 year follow up period). Paired t tests were used to assess differences in individual item scores between visits 1 and 3.

4.4.3.2 *Features*

Features used to train models included 1) all 26 items from the MDS-UPDRS I & II 2) all 33 items within the MDS-UPDRS III.

4.4.3.3 *Comparison of machine learning algorithms*

Using the entire dataset, a variety of classification or regression models (depending on the nature of the prediction) were trained and their accuracies compared across a single repetition of 10 fold cross validation.

4.4.3.4 *Model creation*

Random forests classification or regression models were trained and their accuracies evaluated across 10 repetitions of 10 fold cross validation.

4.4.3.4.1 Evaluation of regression models

The accuracy of regression models was assessed through the calculation of the mean absolute error (the average absolute difference between the true and predicted score). A naïve benchmark, whereby the median of the training data was used as the prediction for the test data, was employed to demonstrate the value of machine learning.

4.4.3.4.2 Evaluation of classification models

The accuracy of classification models was assessed through the calculation of area under the curve values (whereby the data used to assess model accuracy had been excluded from the training process). Where there were more than two classes, the maximum of the posterior probabilities of the two negative classes was subtracted from the posterior probability of the class of interest in order to obtain the input into the Matlab *perfcurve* function.[188]

4.4.3.5 *Model refinement*

4.4.3.5.1 The ranking of items by order of predictor importance

Items were ranked in order of predictor importance and models retrained with an incremental number of items, in order to determine the effect on prediction accuracies.

4.4.3.6 *Predictions*

4.4.3.6.1 MDS-UPDRS III total score

The MDS-UPDRS III total score was derived from the summation of items 1 to 33.

4.4.3.6.2 PD disease severity

Participants were categorized into those with mild (≤32), moderate (≥33 and ≤58) and severe (≥59) PD on the basis of their MDS-UPDRS III score (4.2.2). The AUC for each class (mild/moderate/severe) was calculated in turn.[145]

4.4.3.6.3 Tremor dominant (TD) / postural instability and gait difficulty (PIGD) categorisation

Participants were categorised into those with TD, PIGD and indeterminate groups according to criteria by Stebbins et al.[189] In brief, a Tremor score (summed items: 2.10 (self-reported tremor) and 3.15-18 (researcher assessed rest, postural and kinetic

tremor)) was divided by a PIGD score (summed items: 2.12 (self-reported walking and balance), 2.13 (self-reported freezing), 3.10-3.12 (researcher assessed gait, freezing of gait and postural stability)). TD was the label used where the resulting ratio value was ≥1.5 or where the Tremor score was >0 and the PIGD score was 0. Values ≤1 were awarded the label of PIGD. An indeterminate label was used for values >1.0 and <1.5 or where both the tremor score and the PIGD score were 0.

4.4.3.7 *Creation of abbreviated MDS-UPDRS III score*

Having arranged MDS-UPDRS III items in descending order of importance in the prediction of the total MDS-UPDRS III score, incremental individual MDS-UPDRS III item scores were summed and each resultant score correlated with the total MDS-UPDRS III score.

4.5 Results

4.5.1 Demographics

Baseline characteristics of the 923 participants with idiopathic PD who contributed one or more complete sets of data to the total of 2815 sets, are shown in Figure 21.

Figure 21 | Baseline characteristics of participants contributing one or more complete sets of MDS-UPDRS I-III data during their longitudinal follow up

Age	67.1 (9.6)
Male sex (%)	595 (64%)
Duration since diagnosis (years)	1.2 (0.9)
MDS-UPDRS I[a]	8.8 (5.1)
MDS-UPDRS II[b]	8.7 (6.0)
MDS-UPDRS III[c]	26.5 (10.8)
PD severity (defined according to MDS-UPDRS III[c] score)[145] Mild: Moderate: Severe:	 668 (73%) 236 (26%) 8 (1%)
Hoehn & Yahr[d]	1.8 (0.5)
Education adjusted MoCA[e]	24.9 (3.3)

Values indicate mean (SD) unless otherwise specified. At baseline [a]MDS-UPDRS I values were missing in 15, [b]MDS-UPDRS II values in 10, [c]MDS-UPDRS III values in 11, [d]Hoehn and Yahr values in 3 and [e]MoCA values in 38.

4.5.2 Longitudinal change

502 participants contributed data at baseline, visit 2 (18 months) and visit 3 (36 months). Of the remaining 421 participants, 68 had missed one or more of their first three visits, 188 had withdrawn before completing their third visit, 42 had had one or more follow up assessments by telephone and so did not complete a motor examination (part III) and 129 attended for all three visits but had one or more missing answers to items within one or more MDS-UPDRS parts I-III. Comparing the baseline data of individuals who went on to contribute complete sets of data at visits 2 and 3, with those who did not, individuals who contributed data at all of their first three visits were younger at baseline ($p<0.01$), had a shorter disease duration ($p=0.03$) and lower MDS-UPDRS part I, II and III scores ($p<0.001$). However, there was no difference in sex ($p=0.74$) or TD/PIGD/indeterminate disease subtype ($p=0.59$) between the two groups.

Section 4.5.2 and its subparts describes longitudinal change over the first 3 visits in the 502 participants who contributed complete sets of data at all three visits.

4.5.2.1 *Longitudinal change in MDS-UPDRS I-III*

4.5.2.1.1 Changes in total scores

There were significant ($p<0.001$) differences in total MDS-UPDRS I, II and III scores longitudinally, between visits 1, 2 and 3 (Figure 22, Figure 23 and Figure 24).

Figure 22 | Histograms comparing MDS-UPDRS I score distributions at visits A) 1, B) 2 and C) 3

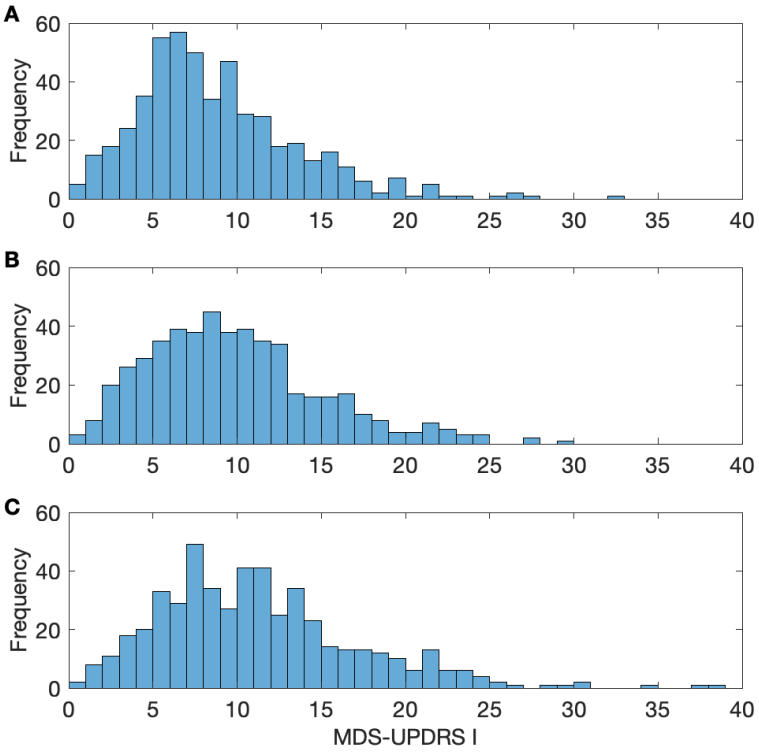


Figure 23 | Histograms comparing MDS-UPDRS II score distributions at visits A) 1, B) 2 and C) 3

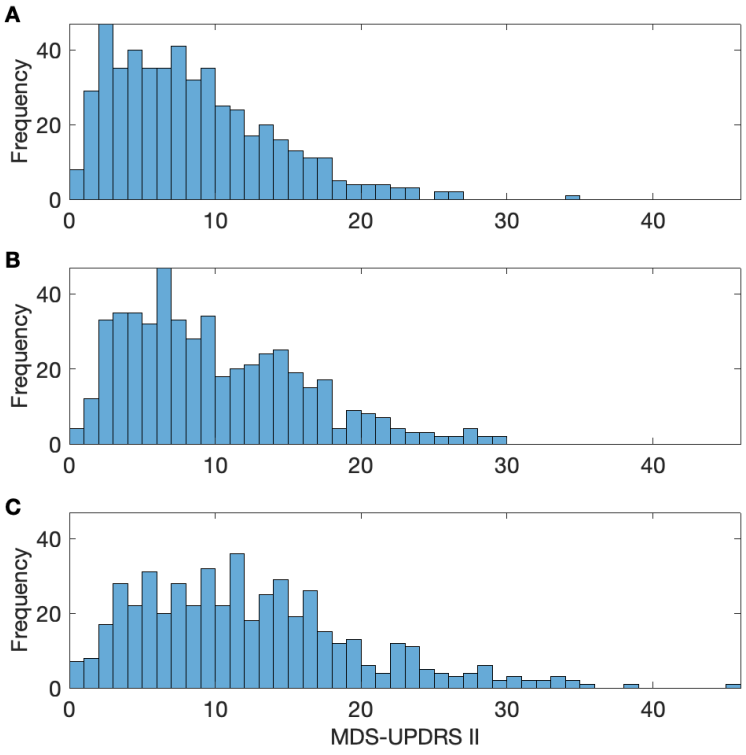
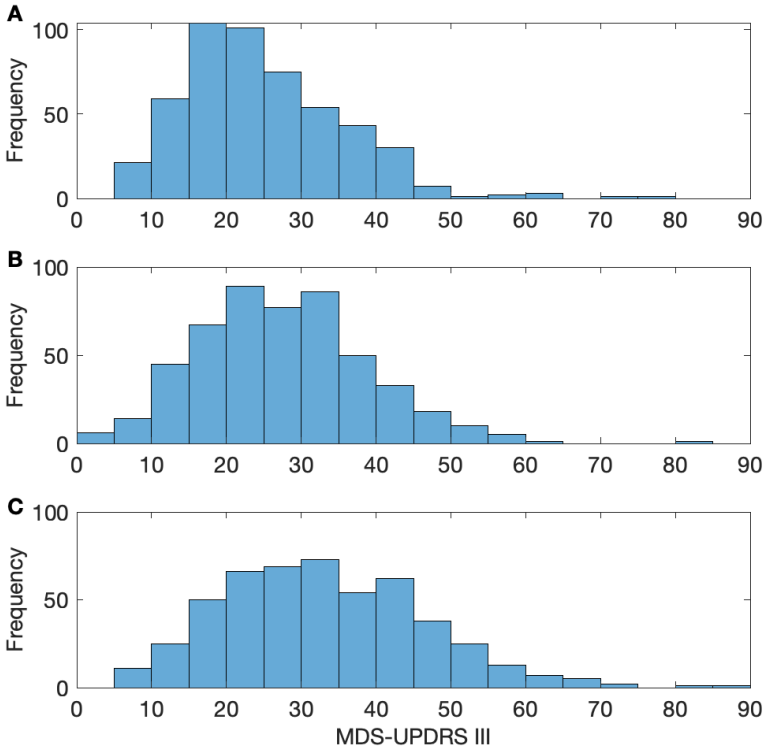


Figure 24 | Histograms comparing MDS-UPDRS III score distributions at visits A) 1, B) 2 and C) 3



4.5.2.1.2 Changes in individual item scores

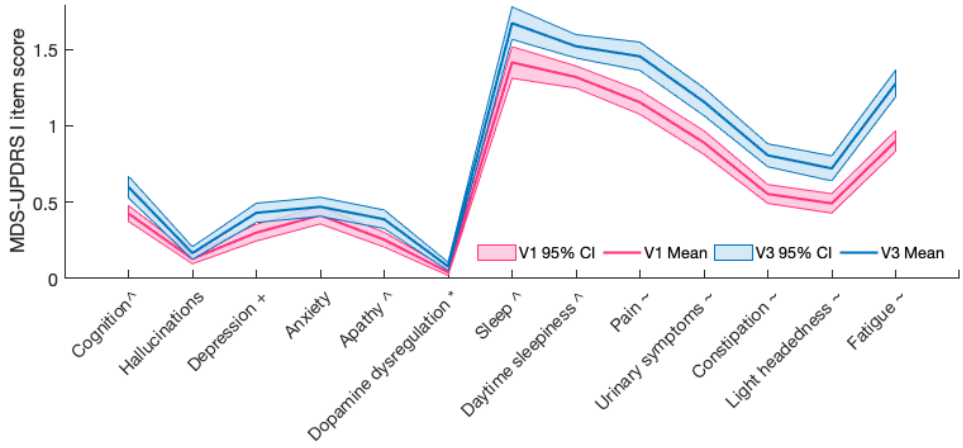
Between visits 1 & 3, within MDS-UPDRS subparts I-III there were statistically significant differences in individual item scores with the exception of the presence of self-reported hallucinations (p=0.16) and anxiety (p=0.19) and researcher examined right hand postural tremor (p=0.36), right (p=0.44) and left (p=0.14) hand kinetic tremor and overall constancy of tremor (p=0.37) (Figure 25).

4.5.2.1.3 Proportion remaining symptom free

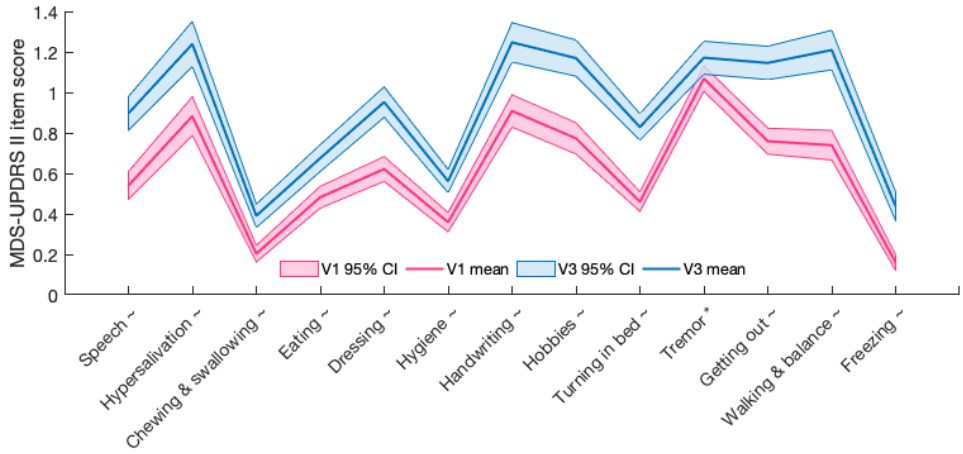
Of the items within MDS-UPDRS parts I-III, researcher assessed freezing, lip and lower limb rest tremor, and researcher assessed hallucinations and features of dopamine dysregulation syndrome were most frequently absent over the 3 year follow up period (Table 18).

Figure 25 | The mean (95% CI) of scores for each item within MDS-UPDRS A) Part I (self-reported non-motor symptom severity) B) Part II (self-reported motor symptom severity) and C) Part III (researcher assessed motor symptom severity)

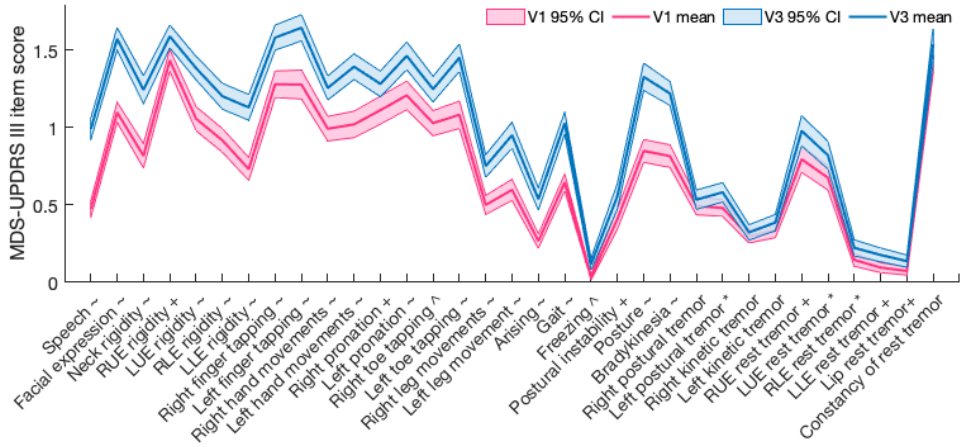
A



B



C



Each item has a maximum score of 4 points; * = p<0.05, + = p<0.01, ^ = p<0.001, ~ = p<0.0001

Table 18 | Proportion of individuals who were symptom free at baseline, 18 months and 3 years follow up for each MDS-UPDRS item within parts I to III

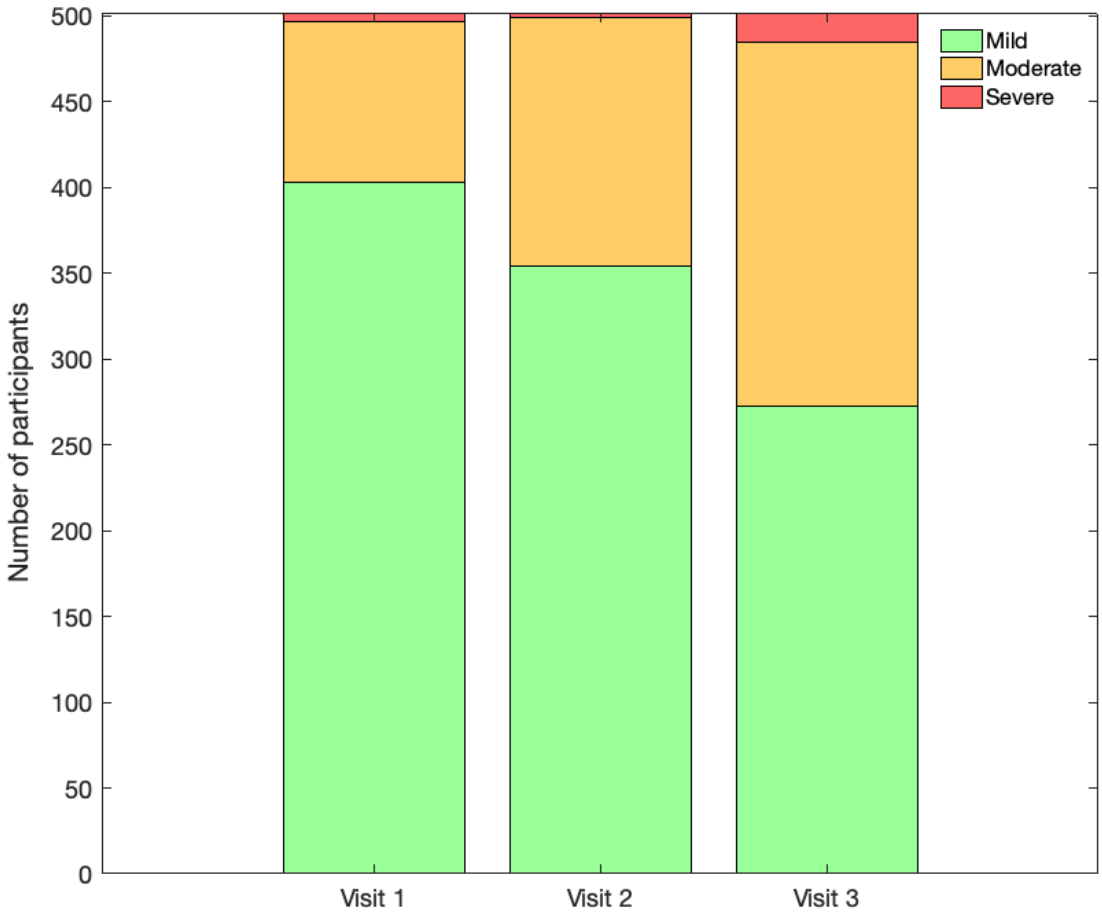
MDS-UPDRS I		% still symptom free after 3 years	MDS-UPDRS II		% still symptom free after 3 years
1.6	Dopamine dysregulation	88	2.13	Freezing	65
1.2	Hallucinations	76	2.3	Chewing & swallowing	57
1.5	Apathy	55	2.6	Hygiene	40
1.3	Depression	51	2.4	Eating	35
1.4	Anxiety	40	2.1	Speech	34
1.1	Cognition	36	2.2	Hypersalivation	32
1.12	Light headedness	35	2.9	Turning in bed	26
1.11	Constipation	28	2.5	Dressing	22
1.10	Urinary symptoms	18	2.7	Handwriting	20
1.7	Sleep	10	2.12	Walking & balance	19
1.13	Fatigue	10	2.8	Hobbies	17
1.8	Daytime sleepiness	7	2.11	Getting out	16
1.9	Pain	5	2.10	Tremor	11

MDS-UPDRS III		% still symptom free after 3 years	MDS-UPDRS III		% still symptom free after 3 years
3.11	Freezing	88	3.17	Constancy of rest tremor	13
3.17	Lip rest tremor	84			
3.17	LLE rest tremor	81	3.10	Gait	11
3.17	RLE rest tremor	75	3.13	Posture	11
3.9	Arising	52	3.3	RLE rigidity	10
3.12	Postural stability	49	3.6	Left pronation	9
3.16	Right kinetic tremor	49	3.5	Right hand movements	9
3.16	Left kinetic tremor	45	3.5	Left hand movements	8
3.17	LUE rest tremor	38	3.6	Right pronation	8
3.15	Right postural tremor	35	3.14	Bradykinesia	8
3.17	RUE rest tremor	34	3.7	Right toe tapping	7
3.15	Left postural tremor	34	3.7	Left toe tapping	7
3.8	Right leg movements	25	3.3	LUE rigidity	6
3.8	Left leg movements	17	3.4	Left finger tapping	5
3.1	Speech	17	3.4	Right finger tapping	5
3.3	LLE rigidity	16	3.2	Facial expression	3
3.3	Neck rigidity	14	3.3	RUE rigidity	2

LLE: Left Lower Extremity, RLE: Right Lower Extremity, LUE: Left Upper Extremity, RUE: Right Upper Extremity

4.5.2.2 *Longitudinal change in MDS-UPDRS disease severity*

Of the 502 participants who contributed complete data at visits 1 to 3, 80% had mild and 19% moderate PD at their first visit, as defined by their MDS-UPDRS III motor examination score.[145] Of the 403 individuals with mild disease at baseline, 60% remained mild at 3 years follow up (their 3rd visit), 38% developed moderate disease and 1% severe disease.



4.5.2.3 *Longitudinal change in TD/PIGD subtype*

The majority (n=380, 76%) of participants who contributed complete data at visits 1 to 3 were classified as TD at baseline. 18% were classified as PIGD and 7% as

indeterminate. At 3 years follow up, 83% of those classified as TD at baseline remained TD; the same was true for 66% of those with baseline PIGD. In general, figures were similar to those reported by the PPMI study[170] at 1 year follow up, the exception being the indeterminate group where a lower percentage (9% versus 36%) of those in the Discovery study remained indeterminate over follow up; the majority converting to PIGD (Table 19).

Table 19 | The proportion of individuals categorised as TD, PIGD and indeterminate at baseline and follow up in the Discovery and PPMI studies.

	Discovery (N=502)	**PPMI (N=320)**
Baseline → follow up interval	3 years	1 year
TD @ baseline	380 (76%)	228 (71%)
TD → TD	314 (83%)	82%
TD → PIGD	48 (13%)	10%
TD → Indeterminate	18 (5%)	8%
PIGD @ baseline	89 (18%)	56 (18%)
PIGD → PIGD	59 (66%)	60%
PIGD → TD	23 (26%)	29%
PIGD → Indeterminate	7 (8%)	11%
Indeterminate @ baseline	33 (7%)	36 (11%)
Indeterminate → Indeterminate	3 (9%)	36%
Indeterminate → TD	12 (36%)	33%
Indeterminate → PIGD	18 (55%)	31%

Values shown within the table represent N(%).

4.5.3 Predictions

4.5.3.1 *MDS-UPDRS III score prediction*

4.5.3.1.1 Using MDS-UPDRS Part I & II items to predict the MDS-UPDRS III

4.5.3.1.1.1 Comparing machine learning approaches, random forests provided close to the lowest mean absolute errors.

Using one repetition of 10 fold cross validation and the raw answers provided for each of the items within MDS-UPDRS parts I & II, random forests regression models were associated with an increase in mean absolute error of 0.2 MDS-UPDRS III points in predicting the MDS-UPDRS III score, over and above that associated with Coarse Gaussian SVM and Exponential Gaussian Process Regression (GPR), associated with the lowest mean absolute errors (Table 20).

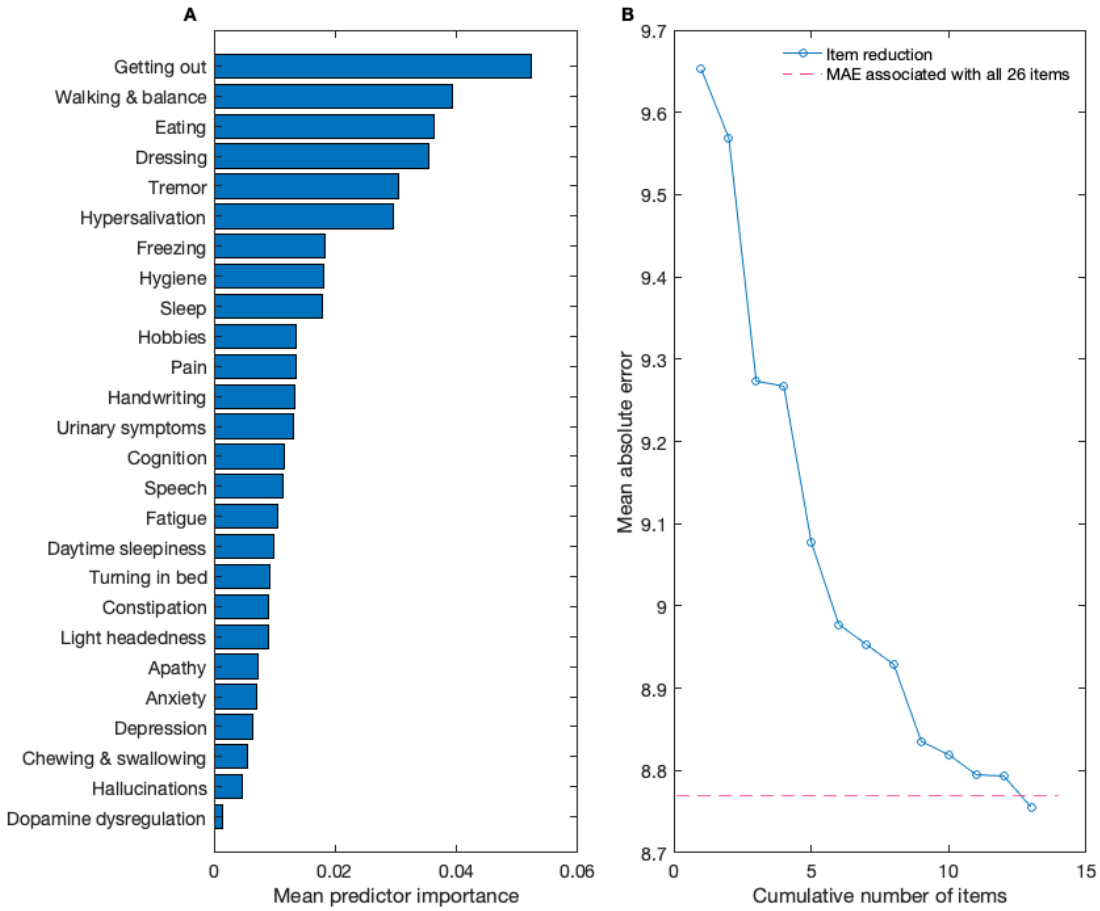
Table 20 | Comparison of regression machine learning approaches in predicting the MDS-UPDRS III score using items within MDS-UPDRS parts I & II

Type	Algorithm	Mean absolute error
Linear Regression	Linear Regression	8.7
	Linear Regression Interactions	9.3
	Linear Regression Robust	8.7
	Stepwise Linear Regression	9
Simple decision trees	Fine Tree	10.7
	Medium Tree	9.8
	Coarse Tree	9.3
Support Vector Machines	Linear SVM	8.7
	Quadratic SVM	9
	Cubic SVM	9.5
	Fine Gaussian SVM	10.2
	Medium Gaussian SVM	8.8
	Coarse Gaussian SVM	8.6
Ensembles	Boosted Trees	8.8
	Random Forests Ensemble	**8.8**
Gaussian Processes Regression	Squared Exponential	8.7
	Matern 5/2	8.7
	Exponential	8.6
	Rational Quadratic	8.7

4.5.3.1.1.2 *Using the 26 items from MDS-UPDRS parts I & II, the total MDS-UPDRS III score was predicted with a mean absolute error (MAE) of 8.8 (SD 0.3) across 10 repetitions of 10 fold cross validation*

Using the top 13 items, by order of predictor importance, the MDS-UPDRS III was predicted with a MAE of 8.8 (SD 0.3) (Figure 26) compared to a naïve error of 10.5 (SD 0.4). Items with the greatest predictor importance tended to reflect motor symptoms. However, self-reported symptoms of dopamine dysregulation demonstrated the lowest value in predicting the MDS-UPDRS III.

Figure 26 | A) The comparative average predictor importance of each item from parts I & II of the MDS-UPDRS in the prediction of the total MDS-UPDRS III score B) The effect of sequential MDS-UPDRS part I & II item selection, by order of predictor importance, on the mean absolute error of MDS-UPDRS III predictions

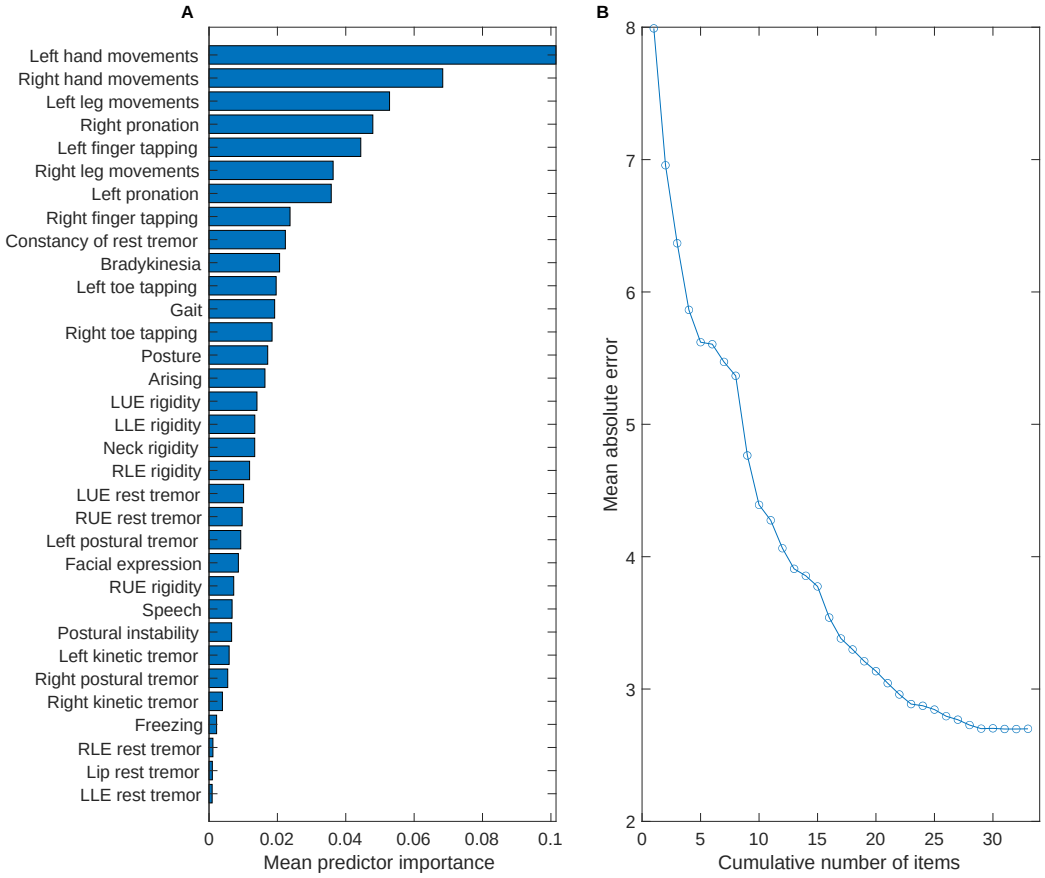


4.5.3.1.2 Using MDS-UPDRS Part III items to predict the MDS-UPDRS III

4.5.3.1.2.1 Using the 33 items from MDS-UPDRS part III, the total MDS-UPDRS III score was predicted with a mean absolute error of 2.7 (SD 0.3) across 10 repetitions of 10 fold cross validation

Due to the nature of random forests (comprising the average prediction of multiple decision trees), the total MDS-UPDRS III score (the sum of the 33 individual item scores) was not predicted exactly. However, training of random forests, allowed the ranking of the 33 items in order of predictor importance. Using the top 10 items, the MDS-UPDRS III was predicted with a MAE of 4.4 (SD 0.2) (Figure 27) compared to a naïve error of 10.5 (SD 0.5) or a MAE of 10.6 (SD 7.8) when linearly scaling up the top 10 item total (out of 40) to the 33 item MDS-UPDRS III total of 132 (i.e. where T_{33} is the 33 item MDS-UPDRS III total and T_{10} is the 10 item score: $T_{33} = T_{10} / 40 \times 132$). Items with the greatest predictor importance tended to involve the assessment of repetitive movements. Lower limb and lip rest tremor were associated with the lowest predictor importance in predicting the MDS-UPDRS III.

Figure 27 | A) The comparative average predictor importance of each item from the MDS-UPDRS part III in predicting its total score B) The effect of sequential MDS-UPDRS III item selection, by order of predictor importance, on the mean absolute error of MDS-UPDRS III total score predictions



4.5.3.2 *PD Disease severity prediction*

Of the 2815 datasets treated as independent, applying previously established cut-off thresholds to the MDS-UPDRS III,[145] 61% were categorised as mild, 36% as moderate and 3% severe.

4.5.3.2.1 Using MDS-UPDRS Part I & II items to predict disease severity

4.5.3.2.1.1 Comparing machine learning approaches, random forests provided close to the highest accuracy values.

Using the entire unbalanced dataset, one repetition of 10 fold cross validation and the raw answers provided for each of the items within MDS-UPDRS parts I & II, there was less than a 1% reduction in random forests classification model accuracy, compared to the highest accuracy values associated with medium gaussian SVM (Table 21).

Table 21 | Comparison of classification machine learning approaches in predicting disease severity using items within MDS-UPDRS parts I & II

Type	Algorithm	Accuracy
Simple decision trees	Fine Tree	64.1%
	Medium Tree	65.4%
	Coarse Tree	64.6%
Discriminant analysis	Linear Discriminant	66.5%
	Quadratic Discriminant	64.6%
Support Vector Machines	Linear SVM	67.0%
	Quadratic SVM	66.2%
	Cubic SVM	63.6%
	Fine Gaussian SVM	61.2%
	Medium Gaussian SVM	67.5%
	Coarse Gaussian SVM	66.8%
Ensembles	Boosted Trees	67.0%
	Random Forests	**66.6%**
	Subspace Discriminant	66.3%
	Subspace KNN	66.8%

4.5.3.2.1.2 *Using the 26 items from MDS-UPDRS parts I & II and 10 fold cross validation, mild, moderate and severe disease was predicted with AUCs of 0.80, 0.62 and 0.86 respectively.*

The same items were ranked in the top 7 and bottom 5, by order of predictor importance, for the prediction of disease severity as well as for the total MDS-UPDRS III score itself (Figure 28). Models trained using the top 9 items and evaluated using 10 fold cross validation yielded classification accuracies comparable with those trained using all 26 items (Figure 29).

Figure 28 | The comparative average predictor importance of each item from parts I & II of the MDS-UPDRS in the prediction of disease severity

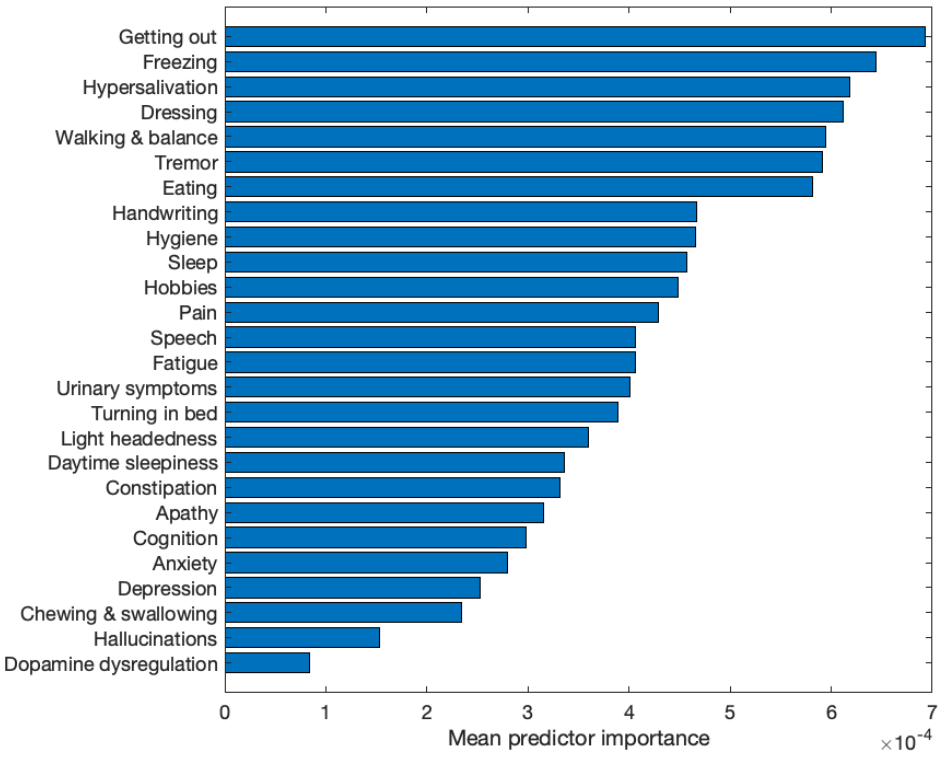
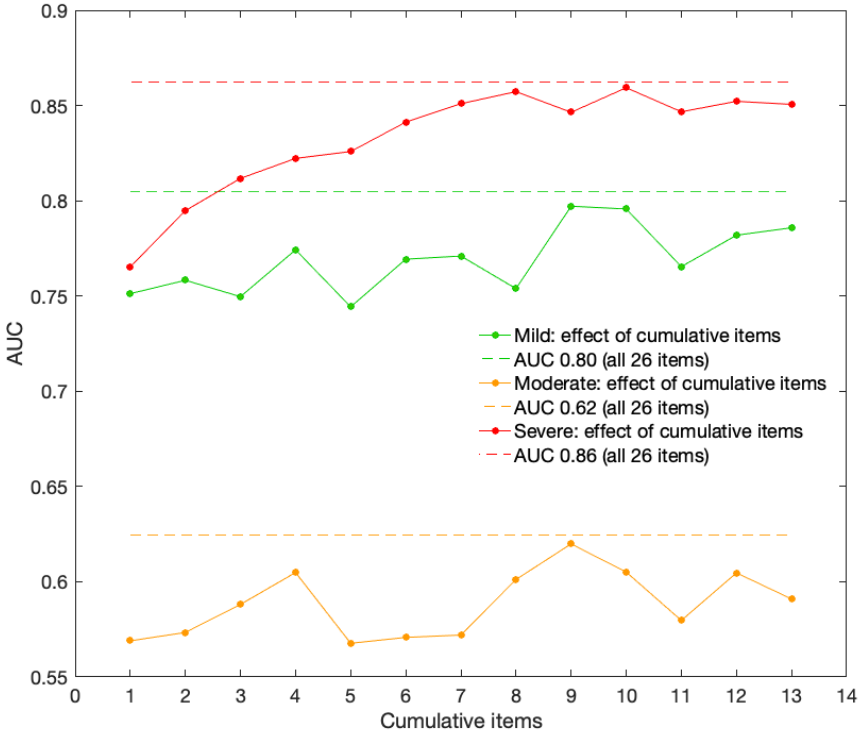


Figure 29 | The effect of sequential MDS-UPDRS I & II item selection, on the accuracy of disease severity classification



4.5.3.2.1.3 Using items from MDS-UPDRS part III and 10 fold cross validation, mild, moderate and severe disease was predicted with AUCs of 0.99, 0.95 and 0.99 respectively.

Unsurprisingly, given the use of the MDS-UPDRS III score in categorizing individuals according to disease severity, there was close homology in the ranking of MDS-UPDRS III items according to predictor importance, in the prediction of both disease severity as well as MDS-UPDRS III score (Figure 27A & Figure 30). Models trained using the top 10 items, each amenable to inertial sensor assessment, were associated with mild, moderate and severe discriminatory AUCs of 0.94, 0.84 and 0.95 respectively (Figure 29). The accuracies associated with the top 16 items, amenable to remote assessment using video conferencing facilities were similarly 0.97, 0.91 and 0.98.

Figure 30 | The comparative average predictor importance of MDS-UPDRS III items in the prediction of disease severity

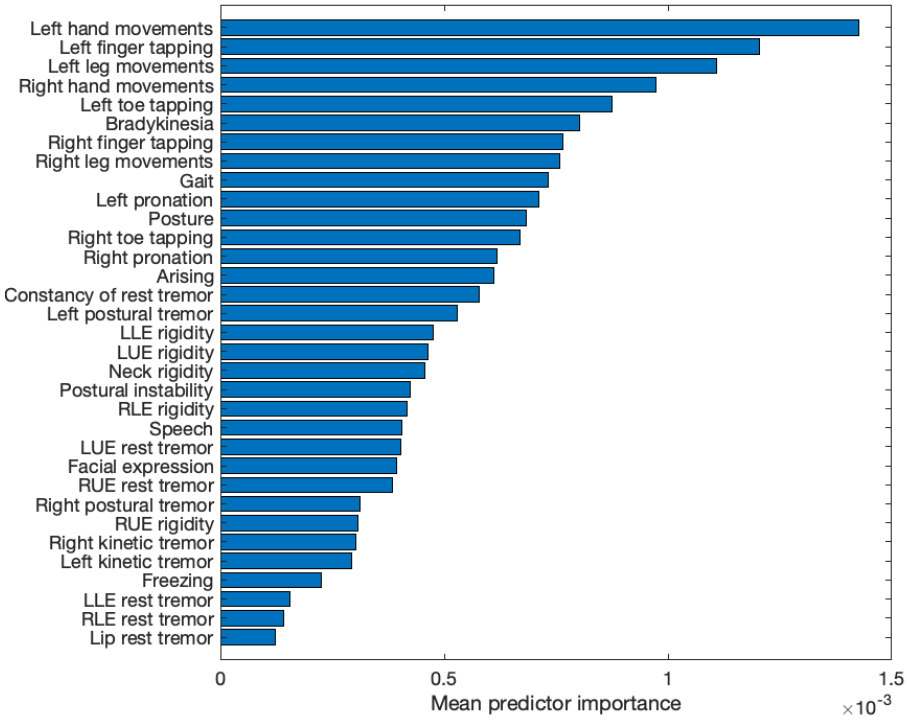
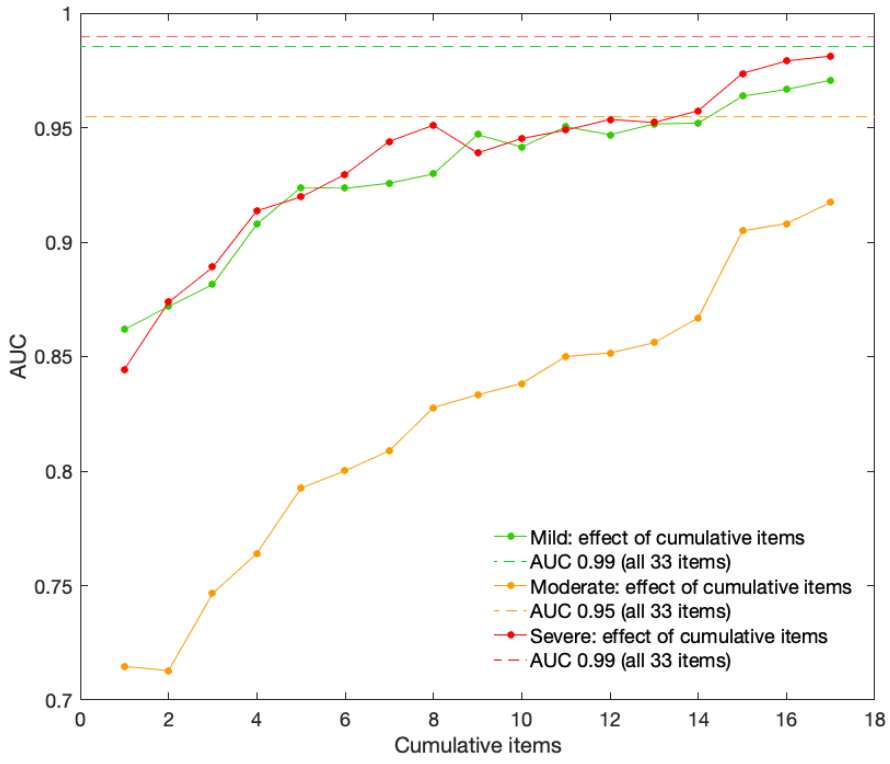


Figure 31 | The effect of sequential MDS-UPDRS III item selection, on the accuracy of disease severity classification



4.5.3.3 *TD/PIGD/indeterminate subtype prediction*

Of the 2815 datasets, treated as independent of each other, 1969 (70%) were categorised as TD, 648 (23%) as PIGD and 198 (7%) as indeterminate using methodology published by Stebbins et al (4.4.3.6.3).[189]

4.5.3.3.1 Comparing machine learning approaches, random forests provided close to the highest accuracy values.

Using the entire unbalanced dataset, one repetition of 10 fold cross validation and the raw answers provided for each of the items within MDS-UPDRS part III, there was less than a 2% reduction in random forests classification model accuracy, compared to the highest accuracy values associated with linear SVM (Table 22).

Table 22 | Comparison of classification machine learning approaches in predicting TD/PIGD/indeterminate subtype using items within MDS-UPDRS part III

Type	Algorithm	Accuracy
Simple decision trees	Fine Tree	84.4%
	Medium Tree	85.4%
	Coarse Tree	83.8%
Discriminant analysis	Linear Discriminant	87.2%
	Quadratic Discriminant	80.9%
Support Vector Machines	Linear SVM	89.0%
	Quadratic SVM	87.7%
	Cubic SVM	87.1%
	Fine Gaussian SVM	70.0%
	Medium Gaussian SVM	88.4%
	Coarse Gaussian SVM	88.8%
k-Nearest Neighbours	Fine KNN	79.2%
	Medium KNN	85.0%
	Coarse KNN	85.1%
	Cosine KNN	84.9%
	Cubic KNN	84.7%
	Weighted KNN	85.1%
Ensembles	Boosted Trees	87.0%
	Random Forests	**87.3%**
	Subspace Discriminant	86.3%
	Subspace KNN	85.8%
	RUSBoosted Trees	79.0%

4.5.3.3.2 Using all items from MDS-UPDRS parts I & II and 10 fold cross validation, TD/PIGD/indeterminate subtypes were predicted with AUCs of 0.80, 0.76 and 0.59 respectively.

Overall, superior accuracy values were obtained by using the answers to self-reported tremor (item 2.10) and walking & balance (item 2.12) alone (AUCs: TD=0.75, PIGD=0.81, Indeterminate=0.64); both constituents in the calculation of the Tremor score:PIGD score ratio (Figure 32 & Figure 33).

Figure 32 | The comparative average predictor importance of MDS-UPDRS I & II items in the prediction of TD/PIGD/indeterminate subtype

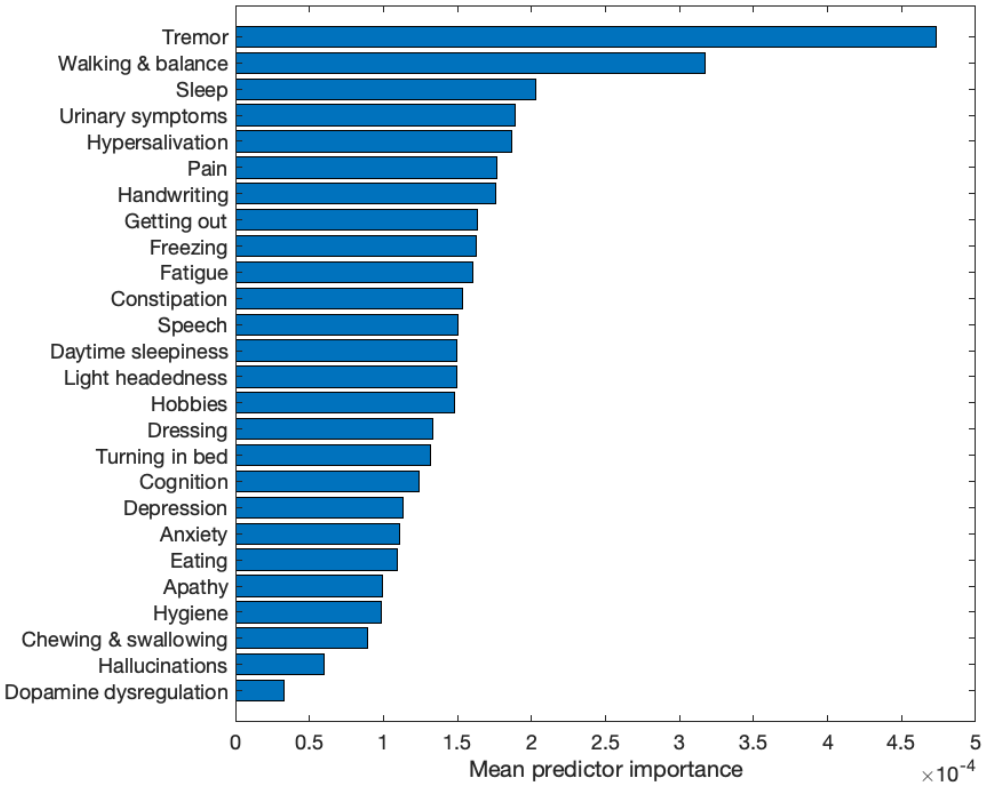
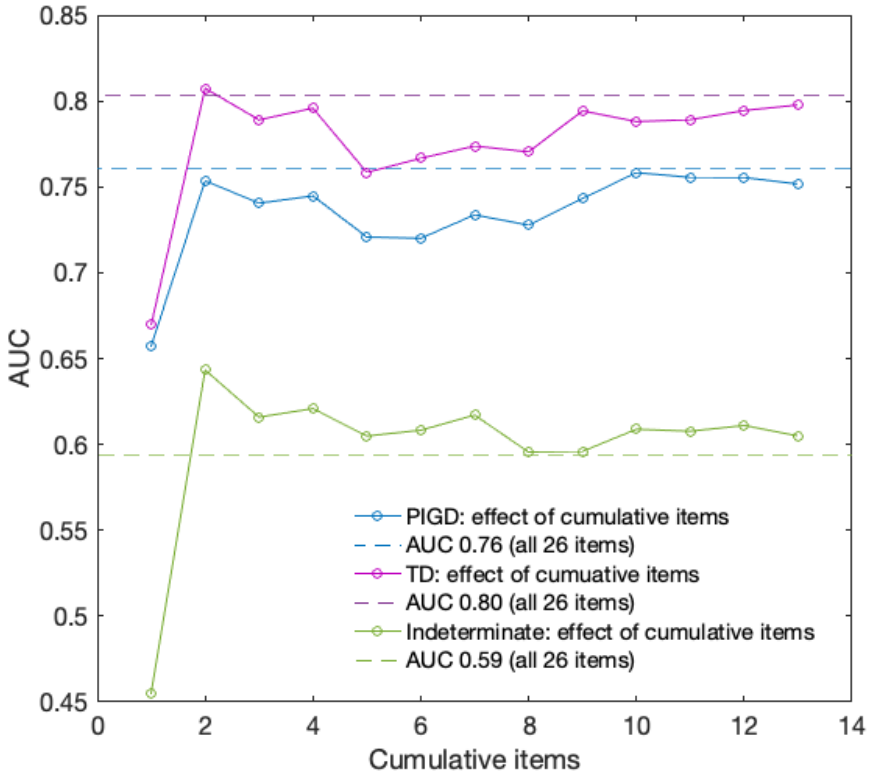


Figure 33 | The effect of sequential MDS-UPDRS I & II item selection, on the accuracy of prediction of TD/PIGD/indeterminate subtype



4.5.3.3.3 Using items from MDS-UPDRS part III and 10 fold cross validation, TD/PIGD/indeterminate subtypes were predicted with AUCs of 0.94, 0.89 and 0.81 respectively (Figure 35).

Comparable accuracies (TD: 0.92, PIGD: 0.89, Indeterminate: 0.80) were achieved using the top 9 items (Figure 34), ranked in order of predictor importance.

Figure 34 | The comparative average predictor importance of MDS-UPDRS III items in the prediction of TD/PIGD/indeterminate subtype

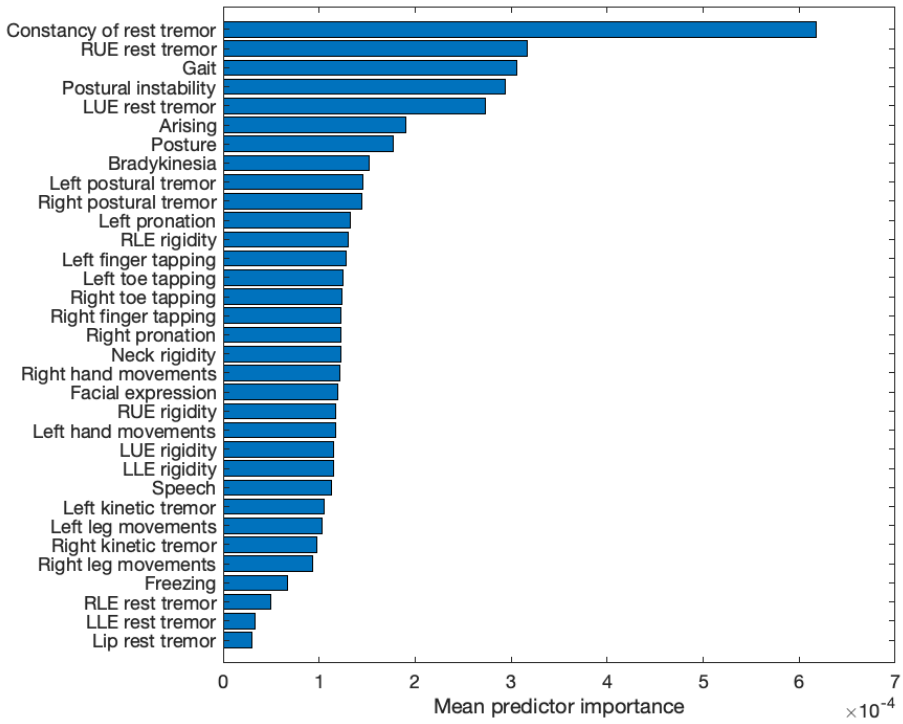
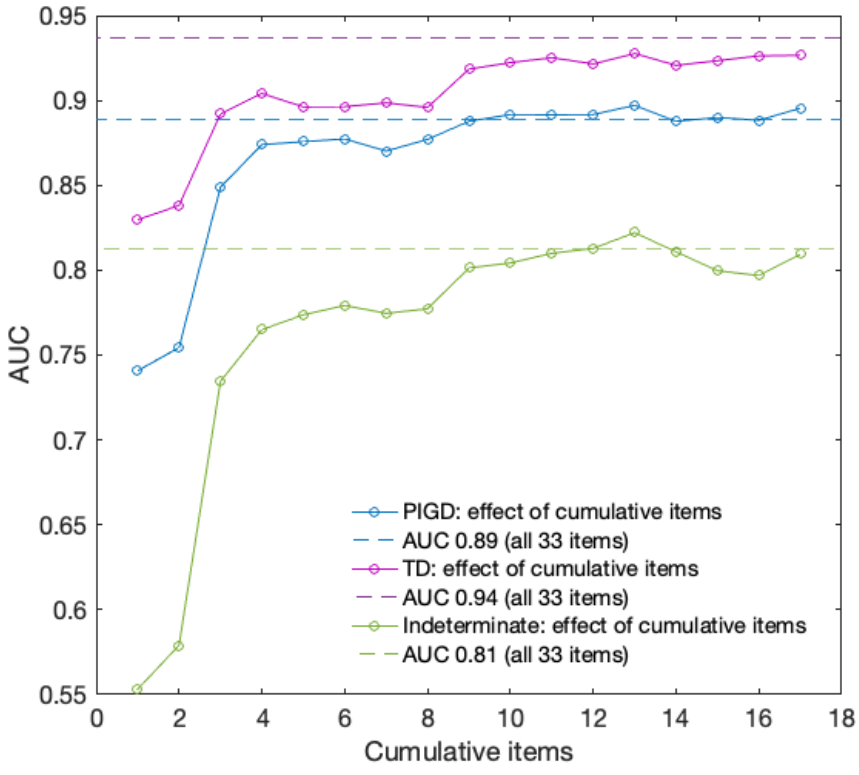


Figure 35 | The effect of sequential MDS-UPDRS III item selection, on the accuracy of prediction of TD/PIGD/indeterminate subtype



4.5.4 Abbreviated MDS-UPDRS III score creation

The top 10 MDS-UPDRS III items, ordered according to their importance in predicting the total MDS-UPDRS III score (Figure 27A), represented 6 tasks that included bilateral assessments of repetitive hand movements, finger tapping, pronation/supination and leg movements, along with overall constancy of rest tremor and bradykinesia. Summation of the top 10 individual item scores, yielded an abbreviated score that had a correlation coefficient of 0.91 ($p<0.0001$) with the MDS-UPDRS III (Figure 36) and a fitted linear polynomial curve was associated with an adjusted r squared of 0.83 (Figure 37).

Figure 36 | Correlation of abbreviated scores with total MDS-UPDRS III

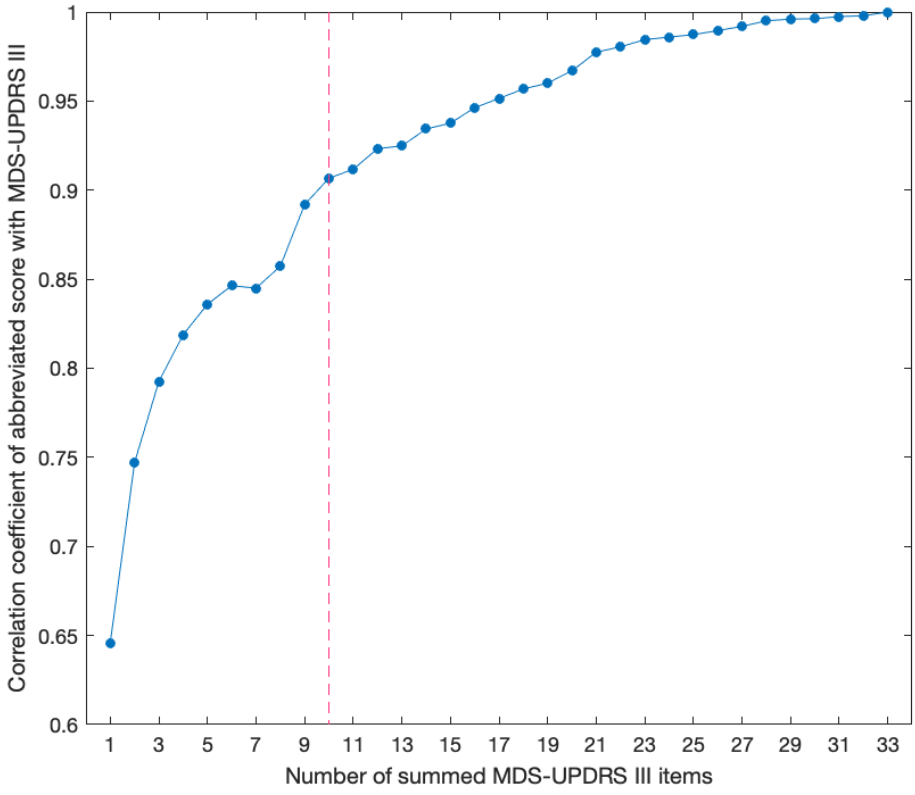
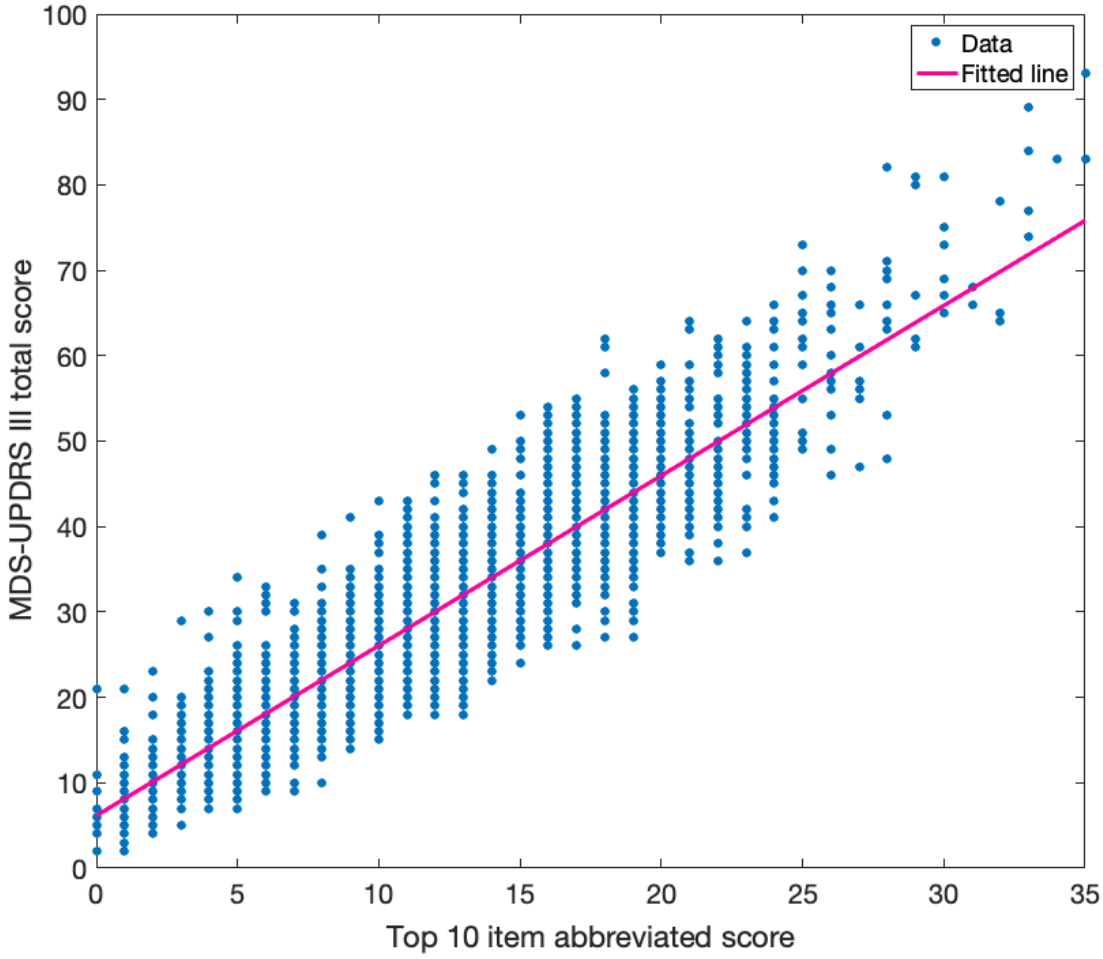


Figure 37 | The top 10 item abbreviated score versus the MDS-UPDRS III total score



Fitted linear model: y = 6.15 (95% 5.70 to 6.60) + 1.99x (95% CI 1.96 to 2.02)

4.6 Discussion

Summary of results:

1. Individual MDS-UPDRS items differ in their longitudinal change
2. Across MDS-UPDRS III score, disease severity & subtype predictions random forests provided close to the highest accuracy values
3. Part I & II items poorly predicted MDS-UPDRS III score, disease severity & subtype
4. Better prediction accuracies were obtained when distinguishing groups at either extreme of disease severity or subtype
5. MDS-UPDRS III items amenable to inertial sensor assessment ranked highly in MDS-UPDRS III score & disease severity prediction
6. An abbreviated MDS-UPDRS III score derived from 10 items demonstrated excellent correlation with the MDS-UPDRS III total score from all 33 items

4.6.1 Individual MDS-UPDRS items differed in their longitudinal change

Whilst MDS-UPDRS subpart I-III scores worsened over time, there were variations in the change in individual item scores, with researcher assessed freezing, lip and lower limb rest tremor and self-reported hallucinations and features of dopamine dysregulation syndrome most likely to be absent over the first 3 visits.

4.6.2 Across MDS-UPDRS III score, disease severity & subtype predictions random forests provided close to the highest accuracy values

In predicting the MDS-UPDRS III score, disease severity and TD/PIGD/indeterminate subtype, there was a negligible difference in accuracy between models trained with random forests and those of SVM models.

4.6.3 Part I & II items poorly predicted MDS-UPDRS III score, disease severity & subtype

Despite the potential advantage of MDS-UPDRS parts I & II being self-administered at home, the MAE in predicting the MDS-UPDRS III score, using all 26 items from both parts I & II (8.8) was greater than the MAE associated with using just the single top-rated item (3.5 Left hand movements) from the MDS-UPDRS part III (7.9). Similarly, the AUCs associated with using all 26 items from MDS-UPDRS I & II were comparable with those obtained by using the top two MDS-UPDRS III items, when predicting disease severity and TD/PIGD/indeterminate subtype.

That within parts I & II, items 2.10 (self-reported tremor) and 2.12 (self-reported walking and balance) were associated with the highest importance in predicting TD/PIGD/indeterminate subtype is unsurprising given the incorporation of their scores into the ratio used to classify individuals according to subtype. Conversely, the ranking of self-reported freezing (item 2.13), below that of non-motor symptoms of sleep, urinary symptoms, hypersalivation, pain, handwriting and difficulty getting out, brings into question the accuracy with which it was reported, given that it too, was used to generate the ratio upon which subtyping was contingent.

4.6.4 Better prediction accuracies were obtained when distinguishing groups at either extreme of disease severity or subtype

The accuracy (AUC) with which PIGD and TD subtypes were distinguished using items from the MDS-UPDRS III was consistently around 0.10 higher than that in distinguishing indeterminate subtypes. A similar difference in AUC values was observed when

distinguishing mild/severe disease compared to that of moderate severity, using the top 10 items.

4.6.5 MDS-UPDRS III items amenable to inertial sensor assessment ranked highly in MDS-UPDRS III score & disease severity prediction

The top 10 ranking of repetitive movements along with gait, overall body bradykinesia and constancy of rest tremor for either prediction 1) highlights the significant contribution of assessments of bradykinesia to the overall grading of motor severity 2) suggests that the constancy of rest tremor is of greater importance than its amplitude 3) points to a lesser role for assessments of rigidity 4) lends credence to the remote assessment of movement using inertial sensors, including those within smartphones as well as dedicated wearable devices.

4.6.6 An abbreviated MDS-UPDRS III score derived from 10 items (6 tasks) demonstrated excellent (rho=0.91) correlation with the MDS-UPDRS III total score from all 33 items

Our results compare favourably with the abbreviated UPDRS-8 score described by Hauser et al., incorporating items from parts I, III and IV of the original UPDRS scale (with one item in common: finger tapping), whose reported correlation with the total UPDRS III score was 0.77.[166] Future work will aim to apply the 10 item abbreviated MDS-UPDRS III score to data from other PD cohorts and to evaluate whether it can be used within a clinical setting (where the scope of motor examinations is often subject to variability between clinicians) to 1) standardise assessments and their documentation and 2) improve treatment decisions.

4.6.7 Limitations

4.6.7.1 *Potential attrition bias*

The dataset analysed in this chapter, from 923 individuals with PD, is one of the largest of its types in the world. As with any longitudinal study, but perhaps more so given the age and co-morbidities of the participants involved, study attrition has the potential to affect results, though bias is by no means inevitable.[190] Statistically significant differences in baseline characteristics were observed between individuals who went on, after their baseline assessment, to contribute complete sets of data and those who did not. The data analysed was largely obtained from individuals with mild and moderate disease. Those of greater disease severity may have been more likely to encounter difficulty completing the study protocol in the presence of cognitive and motor disability. Though an issue shared by many other longitudinal studies involving neurodegenerative diseases, the relative under sampling of individuals with severe disease is still worth being mindful of, when assessing the generalisability of results.

4.6.7.2 *Need for external validation datasets*

Future work will involve the external validation of our findings in independent datasets.

4.6.8 Concluding remarks

This chapter has identified the relative predictor importance of different items within MDS-UPDRS parts I-III, potentially providing evidence that clinicians may use to prioritise aspects of their targeted examination within a busy clinical setting. Rigidity items, which necessitate in-person assessments, were ranked within the middle tertile by order of predictor importance. High MDS-UPDRS III score and disease severity prediction accuracies were obtained using only MDS-UPDRS III items within the upper

tertile of predictor importance, which mainly involved the assessment of repetitive movements, lending credibility to the use of inertial sensor assessments in determining disease severity.

REFERENCES

1. The Incidence and Prevalence of Parkinson's in the UK: Results from the Clinical Practice Research Datalink Reference Report: Parkinson's UK, 2017.

2. Van Den Eeden SK, Tanner CM, Bernstein AL, et al. Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. Am J Epidemiol 2003;157:1015-1022.

3. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology 2018;17:939-953.

4. Dorsey ER, Sherer T, Okun MS, Bloem BR. The Emerging Evidence of the Parkinson Pandemic. Journal of Parkinson's disease 2018;8:S3-s8.

5. Dorsey ER, Bloem BR. The Parkinson Pandemic-A Call to Action. JAMA neurology 2018;75:9-10.

6. Weir S, Samnaliev M, Kuo TC, et al. Short- and long-term cost and utilization of health care resources in Parkinson's disease in the UK. Movement disorders : official journal of the Movement Disorder Society 2018;33:974-981.

7. The cost of Parkinson's: the financial impact of living with the condition: Parkinson's UK, 2017.

8. Parkinson J. An essay on the shaking palsy. London: Sherwood, Neely and Jones 1817.

9. Bernheimer H, Birkmayer W, Hornykiewicz O, Jellinger K, Seitelberger F. Brain dopamine and the syndromes of Parkinson and Huntington. Clinical, morphological and neurochemical correlations. Journal of the neurological sciences 1973;20:415-455.

10. Fearnley JM, Lees AJ. Ageing and Parkinson's disease: substantia nigra regional selectivity. Brain : a journal of neurology 1991;114 (Pt 5):2283-2301.

11. Cheng H-C, Ulane CM, Burke RE. Clinical progression in Parkinson disease and the neurobiology of axons. Annals of neurology 2010;67:715-725.

12. Borghammer P, Van Den Berge N. Brain-First versus Gut-First Parkinson's Disease: A Hypothesis. Journal of Parkinson's disease 2019;9:S281-s295.

13. Braak H, Del Tredici K, Rüb U, de Vos RA, Jansen Steur EN, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiology of aging 2003;24:197-211.

14. Adler CH, Beach TG. Neuropathological basis of nonmotor manifestations of Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2016;31:1114-1119.

15. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 1988;51:745-752.

16. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2015;30:1591-1601.

17. Fereshtehnejad SM, Yao C, Pelletier A, Montplaisir JY, Gagnon JF, Postuma RB. Evolution of prodromal Parkinson's disease and dementia with Lewy bodies: a prospective study. Brain : a journal of neurology 2019;142:2051-2067.

18. Poewe W, Seppi K, Tanner CM, et al. Parkinson disease. Nat Rev Dis Primers 2017;3:17013.

19. Kalia LV, Lang AE. Parkinson's disease. The Lancet 2015;386:896-912.

20. Dickson DW. Neuropathology of Parkinson disease. Parkinsonism Relat Disord 2018;46 Suppl 1:S30-s33.

21. Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. Neurology 2016;86:566-576.

22. Benamer TS, Patterson J, Grosset DG, et al. Accurate differentiation of parkinsonism and essential tremor using visual assessment of [123I]-FP-CIT SPECT imaging: the [123I]-FP-CIT study group. Movement disorders : official journal of the Movement Disorder Society 2000;15:503-510.

23. Marshall VL, Reininger CB, Marquardt M, et al. Parkinson's disease is overdiagnosed clinically at baseline in diagnostically uncertain cases: a 3-year European multicenter study with

repeat [123I]FP-CIT SPECT. Movement disorders : official journal of the Movement Disorder Society 2009;24:500-508.

24. de la Fuente-Fernández R. Role of DaTSCAN and clinical diagnosis in Parkinson disease. Neurology 2012;78:696-701.

25. Shirvan J, Clement N, Ye R, et al. Neuropathologic correlates of amyloid and dopamine transporter imaging in Lewy body disease. Neurology 2019;93:e476-e484.

26. Colloby SJ, McParland S, O'Brien JT, Attems J. Neuropathological correlates of dopaminergic imaging in Alzheimer's disease and Lewy body dementias. Brain : a journal of neurology 2012;135:2798-2808.

27. Ba F, Martin WR. Dopamine transporter imaging as a diagnostic tool for parkinsonism and related disorders in clinical practice. Parkinsonism Relat Disord 2015;21:87-94.

28. Parkinson's disease in adults: NICE guideline, 2017.

29. Athauda D, Foltynie T. The ongoing pursuit of neuroprotective therapies in Parkinson disease. Nature reviews Neurology 2015;11:25-40.

30. Hogl B, Stefani A, Videnovic A. Idiopathic REM sleep behaviour disorder and neurodegeneration - an update. Nature reviews Neurology 2018;14:40-55.

31. Stiasny-Kolster K, Mayer G, Schafer S, Moller JC, Heinzel-Gutenbrunner M, Oertel WH. The REM sleep behavior disorder screening questionnaire--a new diagnostic instrument. Movement disorders : official journal of the Movement Disorder Society 2007;22:2386-2393.

32. Poryazova R, Oberholzer M, Baumann CR, Bassetti CL. REM sleep behavior disorder in Parkinson's disease: a questionnaire-based survey. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine 2013;9:55-59a.

33. Bjørnarå KA, Dietrichs E, Toft M. REM sleep behavior disorder in Parkinson's disease--is there a gender difference? Parkinsonism Relat Disord 2013;19:120-122.

34. Szewczyk-Krolikowski K, Tomlinson P, Nithi K, et al. The influence of age and gender on motor and non-motor features of early Parkinson's disease: initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort. Parkinsonism Relat Disord 2014;20:99-105.

35. Bugalho P, da Silva JA, Neto B. Clinical features associated with REM sleep behavior disorder symptoms in the early stages of Parkinson's disease. Journal of neurology 2011;258:50-55.

36. Bugalho P, Viana-Baptista M. REM sleep behavior disorder and motor dysfunction in Parkinson's disease - A longitudinal study. Parkinsonism & Related Disorders 2013;19:1084-1087.

37. Sixel-Döring F, Zimmermann J, Wegener A, Mollenhauer B, Trenkwalder C. The Evolution of REM Sleep Behavior Disorder in Early Parkinson Disease. Sleep 2016;39:1737-1742.

38. Gagnon JF, Bédard MA, Fantini ML, et al. REM sleep behavior disorder and REM sleep without atonia in Parkinson's disease. Neurology 2002;59:585-589.

39. Postuma RB, Gagnon JF, Vendette M, Charland K, Montplaisir J. REM sleep behaviour disorder in Parkinson's disease is associated with specific motor features. Journal of neurology, neurosurgery, and psychiatry 2008;79:1117-1121.

40. Nomura T, Inoue Y, Kagimura T, Nakashima K. Clinical significance of REM sleep behavior disorder in Parkinson’s disease. Sleep medicine 2013;14:131-135.

41. Sixel-Doring F, Trautmann E, Mollenhauer B, Trenkwalder C. Associated factors for REM sleep behavior disorder in Parkinson disease. Neurology 2011;77:1048-1054.

42. Fereshtehnejad SM, Romenets SR, Anang JB, Latreille V, Gagnon JF, Postuma RB. New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression: A Prospective Cohort Comparison With Other Phenotypes. JAMA neurology 2015;72:863-873.

43. Sasai-Sakuma T, Takeuchi N, Asai Y, Inoue Y, Inoue Y. Prevalence and clinical characteristics of REM sleep behavior disorder in Japanese elderly people. Sleep 2020.

44. Haba-Rubio J, Frauscher B, Marques-Vidal P, et al. Prevalence and Determinants of REM Sleep Behavior Disorder in the General Population. Sleep 2017.

45. Pujol M, Pujol J, Alonso T, et al. Idiopathic REM sleep behavior disorder in the elderly Spanish community: a primary care center study with a two-stage design using video-polysomnography. Sleep medicine 2017;40:116-121.

46. Mahlknecht P, Seppi K, Frauscher B, et al. Probable RBD and association with neurodegenerative disease markers: A population-based study. Movement disorders : official journal of the Movement Disorder Society 2015;30:1417-1421.

47. Boot BP, Boeve BF, Roberts RO, et al. Probable rapid eye movement sleep behavior disorder increases risk for mild cognitive impairment and Parkinson disease: a population-based study. Annals of neurology 2012;71:49-56.

48. Wong JC, Li J, Pavlova M, et al. Risk factors for probable REM sleep behavior disorder: A community-based study. Neurology 2016;86:1306-1312.

49. Li SX, Wing YK, Lam SP, et al. Validation of a new REM sleep behavior disorder questionnaire (RBDQ-HK). Sleep medicine 2010;11:43-48.

50. Boeve BF, Molano JR, Ferman TJ, et al. Validation of the Mayo Sleep Questionnaire to screen for REM sleep behavior disorder in an aging and dementia cohort. Sleep medicine 2011;12:445-453.

51. Frauscher B, Ehrmann L, Zamarian L, et al. Validation of the Innsbruck REM sleep behavior disorder inventory. Movement disorders : official journal of the Movement Disorder Society 2012;27:1673-1678.

52. Boeve BF, Molano JR, Ferman TJ, et al. Validation of the Mayo Sleep Questionnaire to screen for REM sleep behavior disorder in a community-based sample. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine 2013;9:475-480.

53. Fernández-Arcos A, Iranzo A, Serradell M, Gaig C, Santamaria J. The Clinical Phenotype of Idiopathic Rapid Eye Movement Sleep Behavior Disorder at Presentation: A Study in 203 Consecutive Patients. Sleep 2016;39:121-132.

54. Fernández-Arcos A, Morenas-Rodríguez E, Santamaria J, et al. Clinical and video-polysomnographic analysis of rapid eye movement sleep behavior disorder and other sleep disturbances in dementia with Lewy bodies. Sleep 2019;42.

55. Bugalho P, Salavisa M. Factors Influencing the Presentation of REM Sleep Behavior Disorder: The Relative Importance of Sex, Associated Neurological Disorder, and Context of

Referral to Polysomnography. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine 2019;15:1789-1798.

56. Dauvilliers Y, Schenck CH, Postuma RB, et al. REM sleep behaviour disorder. Nature Reviews Disease Primers 2018;4:19.

57. McCarter SJ, Boswell CL, St Louis EK, et al. Treatment outcomes in REM sleep behavior disorder. Sleep medicine 2013;14:237-242.

58. Shin C, Park H, Lee WW, Kim HJ, Kim HJ, Jeon B. Clonazepam for probable REM sleep behavior disorder in Parkinson's disease: A randomized placebo-controlled trial. Journal of the neurological sciences 2019;401:81-86.

59. McGrane IR, Leung JG, St Louis EK, Boeve BF. Melatonin therapy for REM sleep behavior disorder: a critical review of evidence. Sleep medicine 2015;16:19-26.

60. Kunz D, Mahlberg R. A two-part, double-blind, placebo-controlled trial of exogenous melatonin in REM sleep behaviour disorder. Journal of sleep research 2010;19:591-596.

61. Jun JS, Kim R, Byun JI, et al. Prolonged-release melatonin in patients with idiopathic REM sleep behavior disorder. Annals of clinical and translational neurology 2019;6:716-722.

62. Gilat M, Coeytaux Jackson A, Marshall NS, et al. Melatonin for rapid eye movement sleep behavior disorder in Parkinson's disease: A randomised controlled trial. Movement disorders : official journal of the Movement Disorder Society 2020;35:344-349.

63. Barber TR, Lawton M, Rolinski M, et al. Prodromal Parkinsonism and neurodegenerative risk stratification in REM sleep behaviour disorder. Sleep 2017.

64. Postuma RB, Iranzo A, Hu M, et al. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. Brain : a journal of neurology 2019.

65. Heinzel S, Berg D, Gasser T, Chen H, Yao C, Postuma RB. Update of the MDS research criteria for prodromal Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2019.

66. Postuma RB, Gagnon JF, Bertrand JA, Genier Marchand D, Montplaisir JY. Parkinson risk in idiopathic REM sleep behavior disorder: preparing for neuroprotective trials. Neurology 2015;84:1104-1113.

67. Iranzo A, Santamaria J, Tolosa E. Idiopathic rapid eye movement sleep behaviour disorder: diagnosis, management, and the need for neuroprotective interventions. The Lancet Neurology 2016;15:405-419.

68. Schenck CH, Montplaisir JY, Frauscher B, et al. Rapid eye movement sleep behavior disorder: devising controlled active treatment studies for symptomatic and neuroprotective therapy--a consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group. Sleep medicine 2013;14:795-806.

69. McCarter SJ, Sandness DJ, McCarter AR, et al. REM sleep muscle activity in idiopathic REM sleep behavior disorder predicts phenoconversion. Neurology 2019;93:e1171-e1179.

70. Fornazieri MA, de Rezende Pinna F, Bezerra TFP, Antunes MB, Voegels RL. Applicability of the university of pennsylvania smell identification test (SIT) in brazilians: pilot study. Brazilian Journal of Otorhinolaryngology 2010;76:695-699.

71. Wayman C, Hunerlach N. Realising the value of heath care data: a framework for the future: EY, 2019.

72. Post B, Merkus MP, de Bie RM, de Haan RJ, Speelman JD. Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Movement disorders : official journal of the Movement Disorder Society 2005;20:1577-1584.

73. Regnault A, Boroojerdi B, Meunier J, Bani M, Morel T, Cano S. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson’s disease? Learnings from the Parkinson’s progression marker initiative cohort. Journal of neurology 2019;266:1927-1936.

74. UK Communications Market Report: Ofcom, 2018.

75. Espay AJ, Bonato P, Nahab FB, et al. Technology in Parkinson's disease: Challenges and opportunities. Movement disorders : official journal of the Movement Disorder Society 2016;31:1272-1282.

76. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of neurology, neurosurgery, and psychiatry 1992;55:181-184.

77. International Classification of Sleep Disorders. American Academy of Sleep Medicine 2014.

78. Oleszkiewicz A, Schriever VA, Croy I, Hahner A, Hummel T. Updated Sniffin' Sticks normative data based on an extended sample of 9139 subjects. Eur Arch Otorhinolaryngol 2019;276:719-728.

79. Lawton M, Hu MT, Baig F, et al. Equating scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in patients with Parkinson's disease. Parkinsonism Relat Disord 2016;33:96-101.

80. Hummel T, Sekinger B, Wolf SR, Pauli E, Kobal G. 'Sniffin' sticks': olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. Chem Senses 1997;22:39-52.

81. Goetz CG, Tilley BC, Shaftman SR, 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 : official journal of the Movement Disorder Society 2008;23:2129-2170.

82. Li K, Li SH, Su W, Chen HB. Diagnostic accuracy of REM sleep behaviour disorder screening questionnaire: a meta-analysis. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology 2017;38:1039-1046.

83. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53:695-699.

84. Schwab RS, England AC. Projection technique for evaluating surgery in Parkinson’s disease. In: Gillingham FJ, Danoldson IML, ed. Third symposium on Parkinson’s disease; 1969; Edingburgh: Churchill Livingstone: 152-157.

85. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.

86. Jorm AF. A short form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): development and cross-validation. Psychological medicine 1994;24:145-153.

87. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991;39:142-148.

88. Desrosiers J, Hebert R, Bravo G, Dutil E. The Purdue Pegboard Test: normative data for people aged 60 and over. Disabil Rehabil 1995;17:217-224.

89. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology 1967;17:427-442.

90. Giladi N, Shabtai H, Simon ES, Biran S, Tal J, Korczyn AD. Construction of freezing of gait questionnaire for patients with Parkinsonism. Parkinsonism Relat Disord 2000;6:165-170.

91. Breiman L. Random Forests. Machine Learning 2001;45:5-32.

92. Ahmadi Rastegar D, Ho N, Halliday GM, Dzamko N. Parkinson's progression prediction using machine learning and serum cytokines. NPJ Parkinson's disease 2019;5:14.

93. Arora S, Baig F, Lo C, et al. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. Neurology 2018;91:e1528-e1538.

94. Lo C, Arora S, Baig F, et al. Predicting motor, cognitive & functional impairment in Parkinson's. Annals of clinical and translational neurology 2019;6:1498-1509.

95. Ehgoetz Martens KA, Lukasik EL, Georgiades MJ, et al. Predicting the onset of freezing of gait: A longitudinal study. Movement disorders : official journal of the Movement Disorder Society 2018;33:128-135.

96. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological) 1974;36:111-133.

97. Provost F. Machine Learning from Imbalanced Data Sets 1012000.

98. Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering 2014;40:16-28.

99. Doty RL. Olfactory dysfunction in Parkinson disease. Nature reviews Neurology 2012;8:329-339.

100. Baig F, Lawton M, Rolinski M, et al. Delineating nonmotor symptoms in early Parkinson's disease and first-degree relatives. Movement disorders : official journal of the Movement Disorder Society 2015;30:1759-1766.

101. Malek N, Swallow DM, Grosset KA, et al. Olfaction in Parkin single and compound heterozygotes in a cohort of young onset Parkinson's disease patients. Acta Neurol Scand 2016;134:271-276.

102. Kozlovski T, Mitelpunkt A, Thaler A, et al. Hierarchical Data-Driven Analysis of Clinical Symptoms Among Patients With Parkinson's Disease. Frontiers in neurology 2019;10:531.

103. Domellof ME, Lundin KF, Edstrom M, Forsgren L. Olfactory dysfunction and dementia in newly diagnosed patients with Parkinson's disease. Parkinsonism Relat Disord 2017;38:41-47.

104. Fullard ME, Tran B, Xie SX, et al. Olfactory impairment predicts cognitive decline in early Parkinson's disease. Parkinsonism Relat Disord 2016;25:45-51.

105. Schrag A, Siddiqui UF, Anastasiou Z, Weintraub D, Schott JM. Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. The Lancet Neurology 2017;16:66-75.

106. Ham JH, Lee JJ, Sunwoo MK, Hong JY, Sohn YH, Lee PH. Effect of olfactory impairment and white matter hyperintensities on cognition in Parkinson's disease. Parkinsonism Relat Disord 2016;24:95-99.

107. Gjerde KV, Muller B, Skeie GO, Assmus J, Alves G, Tysnes OB. Hyposmia in a simple smell test is associated with accelerated cognitive decline in early Parkinson's disease. Acta Neurol Scand 2018;138:508-514.

108. Kang SH, Lee HM, Seo WK, Kim JH, Koh SB. The combined effect of REM sleep behavior disorder and hyposmia on cognition and motor phenotype in Parkinson's disease. Journal of the neurological sciences 2016;368:374-378.

109. Chen H, Shrestha S, Huang X, et al. Olfaction and incident Parkinson disease in US white and black older adults. Neurology 2017;89:1441-1447.

110. Haehner A, Masala C, Walter S, Reichmann H, Hummel T. Incidence of Parkinson's disease in a large patient cohort with idiopathic smell and taste loss. Journal of neurology 2019;266:339-345.

111. Morley JF, Cheng G, Dubroff JG, Wood S, Wilkinson JR, Duda JE. Olfactory Impairment Predicts Underlying Dopaminergic Deficit in Presumed Drug-Induced Parkinsonism. Mov Disord Clin Pract 2017;4:603-606.

112. Yang HJ, Kim YE, Yun JY, Ehm G, Kim HJ, Jeon BS. Comparison of sleep and other non-motor symptoms between SWEDDs patients and de novo Parkinson's disease patients. Parkinsonism Relat Disord 2014;20:1419-1422.

113. Jennings D, Siderowf A, Stern M, et al. Conversion to Parkinson Disease in the PARS Hyposmic and Dopamine Transporter-Deficit Prodromal Cohort. JAMA neurology 2017;74:933-940.

114. Aguirre-Mardones C, Iranzo A, Vilas D, et al. Prevalence and timeline of nonmotor symptoms in idiopathic rapid eye movement sleep behavior disorder. Journal of neurology 2015;262:1568-1578.

115. Postuma RB, Gagnon JF, Vendette M, Desjardins C, Montplaisir JY. Olfaction and color vision identify impending neurodegeneration in rapid eye movement sleep behavior disorder. Annals of neurology 2011;69:811-818.

116. Chen W, Kang WY, Chen S, et al. Hyposmia correlates with SNCA variant and non-motor symptoms in Chinese patients with Parkinson's disease. Parkinsonism Relat Disord 2015;21:610-614.

117. Shill HA, Hentz JG, Caviness JN, et al. Unawareness of Hyposmia in Elderly People With and Without Parkinson's Disease. Mov Disord Clin Pract 2016;3:43-47.

118. Doty RL, Shaman P, Kimmelman CP, Dann MS. University of Pennsylvania Smell Identification Test: a rapid quantitative olfactory function test for the clinic. Laryngoscope 1984;94:176-178.

119. Morley JF, Cohen A, Silveira-Moriyama L, et al. Optimizing olfactory testing for the diagnosis of Parkinson's disease: item analysis of the university of Pennsylvania smell identification test. NPJ Parkinson's disease 2018;4:2.

120. Huang SF, Chen K, Wu JJ, et al. Odor Identification Test in Idiopathic REM-Behavior Disorder and Parkinson's Disease in China. PloS one 2016;11:e0160199.

121. Campabadal A, Segura B, Junque C, et al. Comparing the accuracy and neuroanatomical correlates of the UPSIT-40 and the Sniffin' Sticks test in REM sleep behavior disorder. Parkinsonism Relat Disord 2019;65:197-202.

122. Casjens S, Eckert A, Woitalla D, et al. Diagnostic value of the impairment of olfaction in Parkinson's disease. PloS one 2013;8:e64735.

123. Mahlknecht P, Pechlaner R, Boesveldt S, et al. Optimizing odor identification testing as quick and accurate diagnostic tool for Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2016;31:1408-1413.

124. Bohnen NI, Gedela S, Kuwabara H, et al. Selective hyposmia and nigrostriatal dopaminergic denervation in Parkinson's disease. Journal of neurology 2007;254:84-90.

125. Mueller C, Renner B. A new procedure for the short screening of olfactory function using five items from the "Sniffin' Sticks" identification test kit. Am J Rhinol 2006;20:113-116.

126. Hummel T, Pfetzing U, Lotsch J. A short olfactory test based on the identification of three odors. Journal of neurology 2010;257:1316-1321.

127. Lawton M, Ben-Shlomo Y, May MT, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. Journal of neurology, neurosurgery, and psychiatry 2018;89:1279-1287.

128. Swallow DM, Lawton MA, Grosset KA, et al. Statins are underused in recent-onset Parkinson's disease with increased vascular risk: findings from the UK Tracking Parkinson's and Oxford Parkinson's Disease Centre (OPDC) discovery cohorts. Journal of neurology, neurosurgery, and psychiatry 2016;87:1183-1190.

129. Kobal G, Klimek L, Wolfensberger M, et al. Multicenter investigation of 1,036 subjects using a standardized method for the assessment of olfactory function combining tests of odor

identification, odor discrimination, and olfactory thresholds. Eur Arch Otorhinolaryngol 2000;257:205-211.

130. Boesveldt S, Verbaan D, Knol DL, et al. A comparative study of odor identification and odor discrimination deficits in Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2008;23:1984-1990.

131. Goetz CG, Stebbins GT, Tilley BC. Calibration of unified Parkinson's disease rating scale scores to Movement Disorder Society-unified Parkinson's disease rating scale scores. Movement Disorders 2012;27:1239-1242.

132. Iranzo A, Serradell M, Vilaseca I, et al. Longitudinal assessment of olfactory function in idiopathic REM sleep behavior disorder. Parkinsonism Relat Disord 2013;19:600-604.

133. Hummel T, Kobal G, Gudziol H, Mackay-Sim A. Normative data for the "Sniffin' Sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects. Eur Arch Otorhinolaryngol 2007;264:237-243.

134. Kawasaki I, Baba T, Takeda A, Mori E. Loss of awareness of hyposmia is associated with mild cognitive impairment in Parkinson's disease. Parkinsonism Relat Disord 2016;22:74-79.

135. Litvan I, Goldman JG, Troster AI, et al. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Movement disorders : official journal of the Movement Disorder Society 2012;27:349-356.

136. Doty RL, Deems DA, Stellar S. Olfactory dysfunction in parkinsonism: a general deficit unrelated to neurologic signs, disease stage, or disease duration. Neurology 1988;38:1237-1244.

137. Herting B, Schulze S, Reichmann H, Haehner A, Hummel T. A longitudinal study of olfactory function in patients with idiopathic Parkinson's disease. Journal of neurology 2008;255:367-370.

138. Hummel T, Konnerth CG, Rosenheim K, Kobal G. Screening of olfactory function with a four-minute odor identification test: reliability, normative data, and investigations in patients with olfactory loss. Ann Otol Rhinol Laryngol 2001;110:976-981.

139. Baig F, Kelly MJ, Lawton MA, et al. Impulse control disorders in Parkinson disease and RBD: A longitudinal study of severity. Neurology 2019;93:e675-e687.

140. Baig F, Lawton MA, Rolinski M, et al. Personality and addictive behaviours in early Parkinson's disease and REM sleep behaviour disorder. Parkinsonism Relat Disord 2017;37:72-78.

141. Rolinski M, Zokaei N, Baig F, et al. Visual short-term memory deficits in REM sleep behaviour disorder mirror those in Parkinson's disease. Brain : a journal of neurology 2016;139:47-53.

142. Fahn S, Elton RL, UPDRS program members. Unified Parkinsons Disease Rating Scale. In: Fahn S, Marsden CD, Goldstein M, Calne DB, editors. Recent developments in Parkinsons disease: Florham Park, NJ: Macmillan Healthcare Information, 1987.

143. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. Movement disorders : official journal of the Movement Disorder Society 2003;18:738-750.

144. Horváth K, Aschermann Z, Ács P, et al. Minimal clinically important difference on the Motor Examination part of MDS-UPDRS. Parkinsonism & Related Disorders 2015;21:1421-1426.

145. Martinez-Martin P, Rodriguez-Blazquez C, Mario A, et al. Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. Parkinsonism Relat Disord 2015;21:50-54.

146. Holden SK, Finseth T, Sillau SH, Berman BD. Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort. Mov Disord Clin Pract 2018;5:47-53.

147. The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol 2011;95:629-635.

148. Makkos A, Kovacs M, Aschermann Z, et al. Are the MDS-UPDRS-Based Composite Scores Clinically Applicable? Movement disorders : official journal of the Movement Disorder Society 2018;33:835-839.

149. Hagell P. Measuring activities of daily living in Parkinson’s disease: On a road to nowhere and back again? Measurement 2019;132:109-124.

150. Siderowf A, McDermott M, Kieburtz K, Blindauer K, Plumb S, Shoulson I. Test-retest reliability of the unified Parkinson's disease rating scale in patients with early Parkinson's disease: results from a multicenter clinical trial. Movement disorders : official journal of the Movement Disorder Society 2002;17:758-763.

151. Richards M, Marder K, Cote L, Mayeux R. Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination. Movement disorders : official journal of the Movement Disorder Society 1994;9:89-91.

152. Evers LJW, Krijthe JH, Meinders MJ, Bloem BR, Heskes TM. Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS. Movement disorders : official journal of the Movement Disorder Society 2019;34:1480-1487.

153. Noyce AJ, Schrag A, Masters JM, Bestwick JP, Giovannoni G, Lees AJ. Subtle motor disturbances in PREDICT-PD participants. Journal of neurology, neurosurgery, and psychiatry 2017;88:212-217.

154. Elkouzi A, Vedam-Mai V, Eisinger RS, Okun MS. Emerging therapies in Parkinson disease - repurposed drugs and new approaches. Nature reviews Neurology 2019;15:204-223.

155. Athauda D, Maclagan K, Skene SS, et al. Exenatide once weekly versus placebo in Parkinson's disease: a randomised, double-blind, placebo-controlled trial. Lancet (London, England) 2017.

156. Initiative CTT. Use case for developing novel endpoints generated using mobile technology: Parkinson’s disease. 2017;1-7. Accessed 23/04/2020.

157. Administration FaD. Clinical Outcome Assessment Compendium. 2015;1-46. Accessed 23/04/200.

158. (CHMP) CfMPfHU. Guideline on clinical investigation of medicinal products in the treatment of Parkinson’s disease. 2012;1-16. Accessed 23/04/2020.

159. Isradipine Versus Placebo in Early Parkinson Disease: A Randomized Trial. Ann Intern Med 2020.

160. Whone A, Luz M, Boca M, et al. Randomized trial of intermittent intraputamenal glial cell line-derived neurotrophic factor in Parkinson's disease. Brain : a journal of neurology 2019;142:512-525.

161. Verschuur CVM, Suwijn SR, Boel JA, et al. Randomized Delayed-Start Trial of Levodopa in Parkinson's Disease. N Engl J Med 2019;380:315-324.

162. Pioglitazone in early Parkinson's disease: a phase 2, multicentre, double-blind, randomised trial. The Lancet Neurology 2015;14:795-803.

163. Beal MF, Oakes D, Shoulson I, et al. A randomized clinical trial of high-dosage coenzyme Q10 in early Parkinson disease: no evidence of benefit. JAMA neurology 2014;71:543-552.

164. Snow BJ, Rolfe FL, Lockhart MM, et al. A double-blind, placebo-controlled study to assess the mitochondria-targeted antioxidant MitoQ as a disease-modifying therapy in Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2010;25:1670-1674.

165. Olanow CW, Rascol O, Hauser R, et al. A double-blind, delayed-start trial of rasagiline in Parkinson's disease. N Engl J Med 2009;361:1268-1278.

166. Hauser RA, Lyons KE, Pahwa R. The UPDRS-8: a brief clinical assessment scale for Parkinson's disease. Int J Neurosci 2012;122:333-337.

167. Abdolahi A, Scoglio N, Killoran A, Dorsey ER, Biglan KM. Potential reliability and validity of a modified version of the Unified Parkinson's Disease Rating Scale that could be administered remotely. Parkinsonism & related disorders 2013;19:218-221.

168. Forjaz MJ, Ayala A, Testa CM, et al. Proposing a Parkinson's disease-specific tremor scale from the MDS-UPDRS. Movement disorders : official journal of the Movement Disorder Society 2015;30:1139-1143.

169. Jankovic J, McDermott M, Carter J, et al. Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. The Parkinson Study Group. Neurology 1990;40:1529-1534.

170. Simuni T, Caspell-Garcia C, Coffey C, Lasch S, Tanner C, Marek K. How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort? Parkinsonism Relat Disord 2016;28:62-67.

171. Herman T, Shema-Shiratzky S, Arie L, Giladi N, Hausdorff JM. Who will remain tremor dominant? The possible role of cognitive reserve in the time course of two common Parkinson's disease motor subtypes. Journal of neural transmission (Vienna, Austria : 1996) 2018;125:1007-1011.

172. Luo L, Andrews H, Alcalay RN, et al. Motor phenotype classification in moderate to advanced PD in BioFIND study. Parkinsonism Relat Disord 2019;65:178-183.

173. Post B, Merkus MP, de Haan RJ, Speelman JD. Prognostic factors for the progression of Parkinson's disease: a systematic review. Movement disorders : official journal of the Movement Disorder Society 2007;22:1839-1851; quiz 1988.

174. van der Heeden JF, Marinus J, Martinez-Martin P, Rodriguez-Blazquez C, Geraedts VJ, van Hilten JJ. Postural instability and gait are associated with severity and prognosis of Parkinson disease. Neurology 2016;86:2243-2250.

175. Rajput AH, Pahwa R, Pahwa P, Rajput A. Prognostic significance of the onset mode in parkinsonism. Neurology 1993;43:829-830.

176. Auyeung M, Tsoi TH, Mok V, et al. Ten year survival and outcomes in a prospective cohort of new onset Chinese Parkinson's disease patients. Journal of neurology, neurosurgery, and psychiatry 2012;83:607-611.

177. Allcock LM, Kenny RA, Burn DJ. Clinical phenotype of subjects with Parkinson's disease and orthostatic hypotension: autonomic symptom and demographic comparison. Movement disorders : official journal of the Movement Disorder Society 2006;21:1851-1855.

178. Rudzińska M, Marona M, Bukowczan S, Banaszkiewicz K, Mirek E, Szczudlik A. Falls in different types of Parkinson's disease. Neurol Neurochir Pol 2007;41:395-403.

179. Hiorth YH, Alves G, Larsen JP, Schulz J, Tysnes OB, Pedersen KF. Long-term risk of falls in an incident Parkinson's disease cohort: the Norwegian ParkWest study. Journal of neurology 2017;264:364-372.

180. Forsaa EB, Larsen JP, Wentzel-Larsen T, Alves G. A 12-year population-based study of freezing of gait in Parkinson's disease. Parkinsonism Relat Disord 2015;21:254-258.

181. Burn DJ, Rowan EN, Minett T, et al. Extrapyramidal features in Parkinson's disease with and without dementia and dementia with Lewy bodies: A cross-sectional comparative study. Movement disorders : official journal of the Movement Disorder Society 2003;18:884-889.

182. Burn DJ, Rowan EN, Allan LM, Molloy S, O'Brien JT, McKeith IG. Motor subtype and cognitive decline in Parkinson's disease, Parkinson's disease with dementia, and dementia with Lewy bodies. Journal of neurology, neurosurgery, and psychiatry 2006;77:585-589.

183. Alves G, Larsen JP, Emre M, Wentzel-Larsen T, Aarsland D. Changes in motor subtype and risk for incident dementia in Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2006;21:1123-1130.

184. Verbaan D, Marinus J, Visser M, et al. Cognitive impairment in Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 2007;78:1182-1187.

185. Aleksovski D, Miljkovic D, Bravi D, Antonini A. Disease progression in Parkinson subtypes: the PPMI dataset. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology 2018;39:1971-1976.

186. Lawton M, Baig F, Rolinski M, et al. Parkinson's Disease Subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery Cohort. Journal of Parkinson's disease 2015;5:269-279.

187. Seidel SE, Tilley BC, Huang P, et al. Subject-investigator reproducibility of the Unified Parkinson's Disease Rating Scale. Parkinsonism Relat Disord 2012;18:230-233.

188. Mathworks. perfcurve documentation [online]. Available at: https://www.mathworks.com/help/stats/perfcurve.html.

189. Stebbins GT, Goetz CG, Burn DJ, Jankovic J, Khoo TK, Tilley BC. How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale. Movement disorders : official journal of the Movement Disorder Society 2013;28:668-670.

190. Lacey RJ, Jordan KP, Croft PR. Does attrition during follow-up of a population cohort study inevitably lead to biased estimates of health status? PloS one 2013;8:e83948-e83948.

191. Greffard S, Verny M, Bonnet AM, et al. Motor score of the Unified Parkinson Disease Rating Scale as a good predictor of Lewy body-associated neuronal loss in the substantia nigra. Archives of neurology 2006;63:584-588.

192. Schrag A, Horsfall L, Walters K, Noyce A, Petersen I. Prediagnostic presentations of Parkinson's disease in primary care: a case-control study. The Lancet Neurology 2015;14:57-64.

193. Tiffin J, Asher EJ. The Purdue pegboard; norms and studies of reliability and validity. J Appl Psychol 1948;32:234-247.

194. Amirjani N, Ashworth NL, Olson JL, Morhart M, Chan KM. Validity and reliability of the Purdue Pegboard Test in carpal tunnel syndrome. Muscle Nerve 2011;43:171-177.

195. Crafton KR, Mark AN, Cramer SC. Improved understanding of cortical injury by incorporating measures of functional anatomy. Brain : a journal of neurology 2003;126:1650-1659.

196. Gallus J, Mathiowetz V. Test-retest reliability of the Purdue Pegboard for persons with multiple sclerosis. Am J Occup Ther 2003;57:108-111.

197. Lee P, Liu CH, Fan CW, Lu CP, Lu WS, Hsieh CL. The test-retest reliability and the minimal detectable change of the Purdue pegboard test in schizophrenia. J Formos Med Assoc 2013;112:332-337.

198. Proud EL, Bilney B, Miller KJ, Morris ME, McGinley JL. Measuring Hand Dexterity in People With Parkinson's Disease: Reliability of Pegboard Tests. Am J Occup Ther 2019;73:7304205050p7304205051-7304205050p7304205058.

199. Ruzicka E, Krupicka R, Zarubova K, Rusz J, Jech R, Szabo Z. Tests of manual dexterity and speed in Parkinson's disease: Not all measure the same. Parkinsonism Relat Disord 2016;28:118-123.

200. Darweesh SK, Wolters FJ, Hofman A, Stricker BH, Koudstaal PJ, Ikram MA. Simple Test of Manual Dexterity Can Help to Identify Persons at High Risk for Neurodegenerative Diseases in the Community. J Gerontol A Biol Sci Med Sci 2017;72:75-81.

201. Soylemez E, Ertugrul S, Dogan E. Assessment of balance skills and falling risk in children with congenital bilateral profound sensorineural hearing loss. Int J Pediatr Otorhinolaryngol 2019;116:75-78.

202. Vancampfort D, Hallgren M, Vandael H, et al. Test-retest reliability and clinical correlates of the Eurofit test battery in people with alcohol use disorders. Psychiatry research 2019;271:208-213.

203. Jacobs JV, Horak FB, Tran VK, Nutt JG. Multiple balance tests improve the assessment of postural stability in subjects with Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 2006;77:322-326.

204. Morris S, Morris ME, Iansek R. Reliability of measurements obtained with the Timed "Up & Go" test in people with Parkinson disease. Phys Ther 2001;81:810-818.

205. van Lummel RC, Walgaard S, Hobert MA, et al. Intra-Rater, Inter-Rater and Test-Retest Reliability of an Instrumented Timed Up and Go (iTUG) Test in Patients with Parkinson's Disease. PloS one 2016;11:e0151881.

206. Haas B, Clarke E, Elver L, Gowman E, Mortimer E, Byrd E. The reliability and validity of the L-test in people with Parkinson's disease. Physiotherapy 2019;105:84-89.

207. Bouca-Machado R, Duarte GS, Patriarca M, et al. Measurement Instruments to Assess Functional Mobility in Parkinson's Disease: A Systematic Review. Mov Disord Clin Pract 2020;7:129-139.

208. Ferrazzoli D, Ortelli P, Riboldazzi G, Maestri R, Frazzitta G. Effectiveness of Rotigotine plus intensive and goal-based rehabilitation versus Rotigotine alone in "de-novo" Parkinsonian subjects: a randomized controlled trial with 18-month follow-up. Journal of neurology 2018;265:906-916.

209. Zhou H, Wang L, Zhang C, et al. Acute Effects of Subthalamic Deep Brain Stimulation on Motor Outcomes in Parkinson's Disease; 13 Year Follow Up. Frontiers in neurology 2019;10:689.

210. Silveira CRA, MacKinley J, Coleman K, et al. Ambroxol as a novel disease-modifying treatment for Parkinson's disease dementia: protocol for a single-centre, randomized, double-blind, placebo-controlled trial. BMC neurology 2019;19:20.

211. Yoo JE, Jang W, Shin DW, et al. Timed Up and Go Test and the Risk of Parkinson's Disease: A Nation-wide Retrospective Cohort Study. Movement disorders : official journal of the Movement Disorder Society 2020.

212. Li Y, Zhang H, Mao W, et al. Visual dysfunction in patients with idiopathic rapid eye movement sleep behavior disorder. Neurosci Lett 2019;709:134360.

213. Postuma RB, Lang AE, Gagnon JF, Pelletier A, Montplaisir JY. How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder. Brain : a journal of neurology 2012;135:1860-1870.

214. Hansen C, Sanchez-Ferro A, Maetzler W. How Mobile Health Technology and Electronic Health Records Will Change Care of Patients with Parkinson's Disease. Journal of Parkinson's disease 2018;8:S41-s45.

215. Erb MK, Karlin DR, Ho BK, et al. mHealth and wearable technology should replace motor diaries to track motor fluctuations in Parkinson's disease. NPJ Digit Med 2020;3:6.

216. Dorsey ER, Vlaanderen FP, Engelen LJ, et al. Moving Parkinson care to the home. Movement disorders : official journal of the Movement Disorder Society 2016;31:1258-1262.

217. Tahafchi P, Judy JW. Freezing-of-Gait Detection Using Wearable Sensor Technology and Possibilistic K-Nearest-Neighbor Algorithm. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2019;2019:4246-4249.

218. Botros A, Schütz N, Camenzind M, et al. Long-Term Home-Monitoring Sensor Technology in Patients with Parkinson's Disease-Acceptance and Adherence. Sensors (Basel, Switzerland) 2019;19:5169.

219. Pardoel S, Kofman J, Nantel J, Lemaire ED. Wearable-Sensor-based Detection and Prediction of Freezing of Gait in Parkinson's Disease: A Review. Sensors (Basel, Switzerland) 2019;19.

220. Ghislieri M, Gastaldi L, Pastorelli S, Tadano S, Agostini V. Wearable Inertial Sensors to Assess Standing Balance: A Systematic Review. Sensors (Basel, Switzerland) 2019;19.

221. Rehman RZU, Del Din S, Shi JQ, et al. Comparison of Walking Protocols and Gait Assessment Systems for Machine Learning-Based Classification of Parkinson's Disease. Sensors (Basel, Switzerland) 2019;19.

222. Bot BM, Suver C, Neto EC, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016;3:160011.

223. Arora S, Baghai-Ravary L, Tsanas A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. J Acoust Soc Am 2019;145:2871.

224. Zhan A, Mohan S, Tarolli C, et al. Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity: The Mobile Parkinson Disease Score. JAMA neurology 2018.

225. Printy BP, Renken LM, Herrmann JP, et al. Smartphone application for classification of motor impairment severity in Parkinson's disease. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2014;2014:2686-2689.

226. Fisher JM, Hammerla NY, Ploetz T, Andras P, Rochester L, Walker RW. Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. Parkinsonism Relat Disord 2016;33:44-50.

227. Arora S, Venkataraman V, Zhan A, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Parkinsonism Relat Disord 2015.

228. Toosizadeh N, Mohler J, Lei H, Parvaneh S, Sherman S, Najafi B. Motor Performance Assessment in Parkinson's Disease: Association between Objective In-Clinic, Objective In-Home, and Subjective/Semi-Objective Measures. PloS one 2015;10:e0124763.

229. Heldman DA, Urrea-Mendoza E, Lovera LC, et al. App-Based Bradykinesia Tasks for Clinic and Home Assessment in Parkinson's Disease: Reliability and Responsiveness. Journal of Parkinson's disease 2017;7:741-747.

230. Del Din S, Godfrey A, Mazza C, Lord S, Rochester L. Free-living monitoring of Parkinson's disease: Lessons from the field. Movement disorders : official journal of the Movement Disorder Society 2016;31:1293-1313.

231. Hasegawa N, Shah VV, Carlson-Kuhta P, Nutt JG, Horak FB, Mancini M. How to Select Balance Measures Sensitive to Parkinson's Disease from Body-Worn Inertial Sensors-Separating the Trees from the Forest. Sensors (Basel, Switzerland) 2019;19.

232. Lee CY, Kang SJ, Hong SK, Ma HI, Lee U, Kim YJ. A Validation Study of a Smartphone-Based Finger Tapping Application for Quantitative Assessment of Bradykinesia in Parkinson's Disease. PloS one 2016;11:e0158852.

233. Zampieri C, Salarian A, Carlson-Kuhta P, Aminian K, Nutt JG, Horak FB. The instrumented timed up and go test: potential outcome measure for disease modifying therapies in Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 2010;81:171-176.

234. Giuffrida JP, Riley DE, Maddux BN, Heldman DA. Clinically deployable Kinesia technology for automated tremor assessment. Movement disorders : official journal of the Movement Disorder Society 2009;24:723-730.

235. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, Kotsavasiloglou C. A Smartphone-Based Tool for Assessing Parkinsonian Hand Tremor. IEEE journal of biomedical and health informatics 2015;19:1835-1842.

236. Arroyo-Gallego T, Ledesma-Carbayo MJ, Sanchez-Ferro A, et al. Detection of Motor Impairment in Parkinson's Disease Via Mobile Touchscreen Typing. IEEE transactions on bio-medical engineering 2017;64:1994-2002.

237. Prince J, Arora S, de Vos M. Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. Physiological measurement 2018;39:044005.

238. Lipsmeier F, Taylor KI, Kilchenmann T, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Movement disorders : official journal of the Movement Disorder Society 2018.

239. Iakovakis D, Hadjidimitriou S, Charisis V, Bostantzopoulou S, Katsarou Z, Hadjileontiadis LJ. Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage Parkinson's disease. Scientific reports 2018;8:7663.

240. Rusz J, Hlavnicka J, Tykalova T, et al. Smartphone Allows Capture of Speech Abnormalities Associated With High Risk of Developing Parkinson's Disease. IEEE Trans Neural Syst Rehabil Eng 2018;26:1495-1507.

241. Mehrang S, Jauhiainen M, Pietil J, Puustinen J, Ruokolainen J, Nieminen H. Identification of Parkinson's Disease Utilizing a Single Self-recorded 20-step Walking Test Acquired by Smartphone's Inertial Measurement Unit. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2018;2018:2913-2916.

242. Prince J, de Vos M. A Deep Learning Framework for the Remote Detection of Parkinson'S Disease Using Smart-Phone Sensor Data. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2018;2018:3144-3147.

243. Lee W, Evans A, Williams DR. Validation of a Smartphone Application Measuring Motor Function in Parkinson's Disease. Journal of Parkinson's disease 2016;6:371-382.

244. Aghanavesi S, Memedi M, Dougherty M, Nyholm D, Westin J. Verification of a Method for Measuring Parkinson's Disease Related Temporal Irregularity in Spiral Drawings. Sensors (Basel, Switzerland) 2017;17.

245. Lee W, Evans AH, Williams DR. Objective Measurement and Characterization of Sleep Benefit in Parkinson's Disease. Mov Disord Clin Pract 2017;4:590-596.

246. Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, Kotsavasiloglou C. Smartphone-based evaluation of parkinsonian hand tremor: quantitative measurements vs clinical assessment scores. Conference proceedings : Annual International Conference of the IEEE Engineering in

Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2014;2014:906-909.

247. Araujo R, Tabuas-Pereira M, Almendra L, et al. Tremor Frequency Assessment by iPhone(R) Applications: Correlation with EMG Analysis. Journal of Parkinson's disease 2016;6:717-721.

248. Bazgir O, Habibi SAH, Palma L, Pierleoni P, Nafees S. A Classification System for Assessment and Home Monitoring of Tremor in Patients with Parkinson's Disease. J Med Signals Sens 2018;8:65-72.

249. Kassavetis P, Saifee TA, Roussos G, et al. Developing a Tool for Remote Digital Assessment of Parkinson's Disease. Mov Disord Clin Pract 2016;3:59-64.

250. Barrantes S, Sanchez Egea AJ, Gonzalez Rojas HA, et al. Differential diagnosis between Parkinson's disease and essential tremor using the smartphone's accelerometer. PloS one 2017;12:e0183843.

251. Yahalom G, Yekutieli Z, Israeli-Korn S, et al. Smartphone Based Timed Up and Go Test Can Identify Postural Instability in Parkinson's Disease. Isr Med Assoc J 2020;22:37-42.

252. Kim H, Lee HJ, Lee W, et al. Unconstrained detection of freezing of Gait in Parkinson's disease patients using smartphone. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2015;2015:3751-3754.

253. Capecci M, Pepa L, Verdini F, Ceravolo MG. A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. Gait Posture 2016;50:28-33.

254. Kim HB, Lee HJ, Lee WW, et al. Validation of Freezing-of-Gait Monitoring Using Smartphone. Telemed J E Health 2018;24:899-907.

255. Ginis P, Nieuwboer A, Dorfman M, et al. Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: A pilot randomized controlled trial. Parkinsonism Relat Disord 2016;22:28-34.

256. Lopez WO, Higuera CA, Fonoff ET, Souza Cde O, Albicker U, Martinez JA. Listenmee and Listenmee smartphone application: synchronizing walking to rhythmic auditory cues to improve gait in Parkinson's disease. Hum Mov Sci 2014;37:147-156.

257. Ivkovic V, Fisher S, Paloski WH. Smartphone-based tactile cueing improves motor performance in Parkinson's disease. Parkinsonism Relat Disord 2016;22:42-47.

258. Fung A, Lai EC, Lee BC. Usability and Validation of the Smarter Balance System: An Unsupervised Dynamic Balance Exercises System for Individuals With Parkinson's Disease. IEEE Trans Neural Syst Rehabil Eng 2018;26:798-806.

259. Tsiouris KM, Gatsios D, Rigas G, et al. PD_Manager: an mHealth platform for Parkinson's disease patient management. Healthcare technology letters 2017;4:102-108.

260. Lopane G, Mellone S, Corzani M, et al. Supervised versus unsupervised technology-based levodopa monitoring in Parkinson's disease: an intrasubject comparison. Journal of neurology 2018;265:1343-1352.

261. Maetzler W, Liepelt I, Berg D. Progression of Parkinson's disease in the clinical phase: potential markers. The Lancet Neurology 2009;8:1158-1171.

262. Velseboer DC, Broeders M, Post B, et al. Prognostic factors of motor impairment, disability, and quality of life in newly diagnosed PD. Neurology 2013;80:627-633.

263. Lawton M, Ben-Shlomo Y, Baig F, et al. Developing and validating Parkinson’s Disease Subtypes and their motor and cognitive progression. JNNP 2018.

264. Bloxham CA, Dick DJ, Moore M. Reaction times and attention in Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 1987;50:1178-1183.

265. Lin HY, Hwang-Gu SL, Gau SS. Intra-individual reaction time variability based on ex-Gaussian distribution as a potential endophenotype for attention-deficit/hyperactivity disorder. Acta psychiatrica Scandinavica 2015;132:39-50.

266. Sejdic E, Lowry KA, Bellanca J, Redfern MS, Brach JS. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. IEEE Trans Neural Syst Rehabil Eng 2014;22:603-612.

267. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology 2017;186:1026-1034.

268. Pagano G, Niccolini F, Politis M. Imaging in Parkinson's disease. Clin Med (Lond) 2016;16:371-375.

269. Iranzo A, Lomeña F, Stockner H, et al. Decreased striatal dopamine transporter uptake and substantia nigra hyperechogenicity as risk markers of synucleinopathy in patients with idiopathic rapid-eye-movement sleep behaviour disorder: a prospective study [corrected]. The Lancet Neurology 2010;9:1070-1077.

270. Iranzo A, Valldeoriola F, Lomeña F, et al. Serial dopamine transporter imaging of nigrostriatal function in patients with idiopathic rapid-eye-movement sleep behaviour disorder: a prospective study. The Lancet Neurology 2011;10:797-805.

271. Hely MA, Reid WG, Adena MA, Halliday GM, Morris JG. The Sydney multicenter study of Parkinson's disease: the inevitability of dementia at 20 years. Movement disorders : official journal of the Movement Disorder Society 2008;23:837-844.

272. Hiorth YH, Larsen JP, Lode K, Pedersen KF. Natural history of falls in a population-based cohort of patients with Parkinson's disease: an 8-year prospective study. Parkinsonism Relat Disord 2014;20:1059-1064.

273. Okuma Y, Silva de Lima AL, Fukae J, Bloem BR, Snijders AH. A prospective study of falls in relation to freezing of gait and response fluctuations in Parkinson's disease. Parkinsonism Relat Disord 2018;46:30-35.

274. Kempster PA, Williams DR, Selikhova M, Holton J, Revesz T, Lees AJ. Patterns of levodopa response in Parkinson's disease: a clinico-pathological study. Brain : a journal of neurology 2007;130:2123-2128.

275. Williams-Gray CH, Mason SL, Evans JR, et al. The CamPaIGN study of Parkinson's disease: 10-year outlook in an incident population-based cohort. Journal of neurology, neurosurgery, and psychiatry 2013;84:1258-1264.

276. Paul SS, Canning CG, Sherrington C, Lord SR, Close JC, Fung VS. Three simple clinical tests to accurately predict falls in people with Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2013;28:655-662.

277. Duncan RP, Cavanaugh JT, Earhart GM, et al. External validation of a simple clinical tool used to predict falls in people with Parkinson disease. Parkinsonism Relat Disord 2015;21:960-963.

278. Velseboer DC, de Bie RM, Wieske L, et al. Development and external validation of a prognostic model in newly diagnosed Parkinson disease. Neurology 2016;86:986-993.

279. Hu MT, Szewczyk-Krolikowski K, Tomlinson P, et al. Predictors of cognitive impairment in an early stage Parkinson's disease cohort. Movement disorders : official journal of the Movement Disorder Society 2014;29:351-359.

280. Giladi N, Tal J, Azulay T, et al. Validation of the freezing of gait questionnaire in patients with Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2009;24:655-661.

281. Hoops S, Nazem S, Siderowf AD, et al. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. Neurology 2009;73:1738-1745.

282. Arora S, Baig F, Lo C, et al. Smartphone motor testing to distinguish idiopathic REM sleep behaviour disorder, controls and PD. Neurology 2018.

283. Harrison JK, Stott DJ, McShane R, Noel-Storr AH, Swann-Price RS, Quinn TJ. Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the early diagnosis of dementia across a variety of healthcare settings. Cochrane Database Syst Rev 2016;11:CD011333.

284. Chahine LM, Iranzo A, Fernandez-Arcos A, et al. Basic clinical features do not predict dopamine transporter binding in idiopathic REM behavior disorder. NPJ Parkinson's disease 2019;5:2.

285. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta psychiatrica Scandinavica 1983;67:361-370.

286. Fasano A, Canning CG, Hausdorff JM, Lord S, Rochester L. Falls in Parkinson's disease: A complex and evolving picture. Movement disorders : official journal of the Movement Disorder Society 2017;32:1524-1536.

287. Walton CC, Mowszowski L, Gilat M, et al. Cognitive training for freezing of gait in Parkinson's disease: a randomized controlled trial. NPJ Parkinson's disease 2018;4:15.

288. Perez-Lloret S, Negre-Pages L, Damier P, et al. Prevalence, determinants, and effect on quality of life of freezing of gait in parkinson disease. JAMA neurology 2014;71:884-890.

289. Liu R, Umbach DM, Troster AI, Huang X, Chen H. Non-motor symptoms and striatal dopamine transporter binding in early Parkinson's disease. Parkinsonism Relat Disord 2020;72:23-30.

290. Lawson RA, Yarnall AJ, Duncan GW, et al. Stability of mild cognitive impairment in newly diagnosed Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 2017;88:648-652.

291. Human Development Report: United Nations Development Programme, 2019.

292. Reilmann R, Rouzade-Dominguez M-L, Saft C, et al. A randomized, placebo-controlled trial of AFQ056 for the treatment of chorea in Huntington's disease. Movement Disorders 2015;30:427-431.