Prediction and real-life monitoring of DBS motor response in Parkinson's disease

Prediction and real-life monitoring of DBS motor response in Parkinson's disease

ISBN: 978-94-6423-412-1
Layout: Jeroen Habets en Ipskamp Printing
Cover design: Perenboom, Van Ham en Groot-Bruinderink Design (‘Vooraluwkaften’©)
Production: Ipskamp Printing | Universitaire Pers Maastricht

The research described in this was conducted at the department of Neurosurgery of Maastricht University, the Maastricht University Medical Center, and the School of Mental Health and Neurosciences, located in Maastricht, The Netherlands.

This research received funding from the Weijerhorst Foundation (received by prof. Y. Temel), and from personal grants rewarded to Jeroen Habets by the Young European Research Universities Network, the Dutch Parkinson Association and the Dutch Research Counsil (ZonMW).

Prediction and real-life monitoring of DBS motor response in Parkinson's disease

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. dr. Rianne M. Letschert
volgens het besluit van het College van Decanen,
In het openbaar te verdedigen op 1 Oktober 2021, om 10.00 uur

door

Jeroen Godefridus Victor Habets

Promotor
Prof. dr. Y. Temel

Copromotores
Dr. P. Kubben
Dr. M. Kuijf

Beoordelingscommissie
Prof. dr. W. Mess (voorzitter)
Dr. P. Bonizzi
Prof. dr. P. Delespaul
Dr. W.J. Neumann (Charité Universitätsmedizin Berlin)
Prof. dr. V. Visser Vandewalle (Universitätsklinikum Köln)

Table of contents

Chapter 1

Introduction

The multidisciplinary topic of this thesis

Since deep brain stimulation (DBS) has been introduced as a 'last resort' treatment for advanced Parkinson's disease (PD) patients in the late 1980s[1], it has become an accepted and successful treatment for both severe and early motor symptoms[2–5]. Despite advances in scientific and clinical knowledge, the current DBS outcome for PD leaves room for improvement: 1) not all PD patients experience a satisfying motor response [3,6], and 2) not all patients show a constant satisfying balance between beneficial motor response and adverse effects7,8.

The work in this thesis aims to improve the current DBS practice for PD in two different ways. In part A, we will discuss the development and validation of a novel preoperative prediction methodology that identifies which PD DBS-candidates are at risk of a not-satisfying motor response. In part B, we aim to improve motion sensor-based PD monitoring. In particular, we focus on monitor validity in patients' real-life environment, with specific attention for characteristics required for motion sensor-based DBS-optimization.

The multidisciplinary work in this thesis combines neurosurgical and neurological experience, with knowledge and skills from computational sciences and clinical data science (figure 1). This multidisciplinary approach is both promising and challenging at the same time. Promising, because technological and computational advances, often generalized as artificial intelligence (AI), offer new possibilities to collect and analyze data, and therefore to answer clinical neurological questions[9]. Challenging, because only a small part of AI-research is actually implemented in clinical practice. Mateen et al suggest the latter can be overcome by conducting multidisciplinary research following recent guidelines[10], and by considering the full translational circle from clinical question, via computational data-driven problem-solving, to utilization in clinical practice[11].

The next paragraphs of this introduction will provide the reader with a brief, but nonetheless complete, introduction of the multidisciplinary knowledge required to interpret the scientific content of this thesis. Paragraph 2 provides clinical knowledge on the etiology, symptomatology, symptom monitoring, and therapy of Parkinson's disease. Paragraph 3 focusses specifically on motion sensor monitoring for Parkinson monitoring. Basic clinical data science concepts used in both parts of the thesis are introduced in paragraph 4. The problem statements of this thesis are outlined per chapter in paragraph 5.

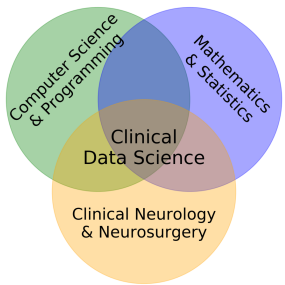


Figure 1: Definition of clinical data science according to the Conway Venn diagram. Research which combines the three visualized skills and competences can be defined as clinical data science. The clinical component is specified for the work in this thesis.

1. Parkinson's disease

1.1 Etiology, anatomy, and pathophysiology

The clinical manifestation currently known as Parkinson's disease (PD) was first described in 1817 by James Parkinson in his 'Essay of the shaking palsy' [12]. Two centuries later, PD is the second

most common neurodegenerative disease after dementia [13]. A global increase of life expectancy contributes to PD's growing incidence, and 12 Million people are expected to be affected worldwide by 2040[13].

In general, PD is regarded as a disease of the basal ganglia. The basal ganglia are a collection of brain regions deep in the brain involved in motor and cognitive functioning, and they consist of the globus pallidum, caudate nucleus, putamen, substantia nigra, and the subthalamic nucleus (STN)[14]. In a healthy brain, the basal ganglia facilitate successful initiation, execution, and control of movement via three motor pathways, the direct, indirect, and hyperdirect pathway (figure 2, left panel, A and B). Basal ganglia regions process and smoothen input signals from higher located cortical regions, before they project their output on other basal ganglia regions, or the thalamus. The thalamus acts as a large 'regulation center' for motor and sensory signals, as well as alertness. Cortical motor signals are processed by the thalamus and projected back to the motor cortex, to travel then via the spinal cord towards peripheral nerves, to eventually activate limb muscles.
For its cognitive functions, the basal ganglia are connected with the frontal and prefrontal cortex of the brain[15]. These cortical regions are involved in higher cognitive functions such as executive functions, planning, memory, impulse control, social and sexual behavior.
To function properly, the connections in the basal ganglia need specific neurotransmitters to transmit signals between connected brain cells, so called neurons. Neurotransmitters are small particles excreted by specific neurons, such as dopamine, and stimulate (excite), or suppress (inhibit) the 'synaptic' signal transduction between two communicating brain cells.

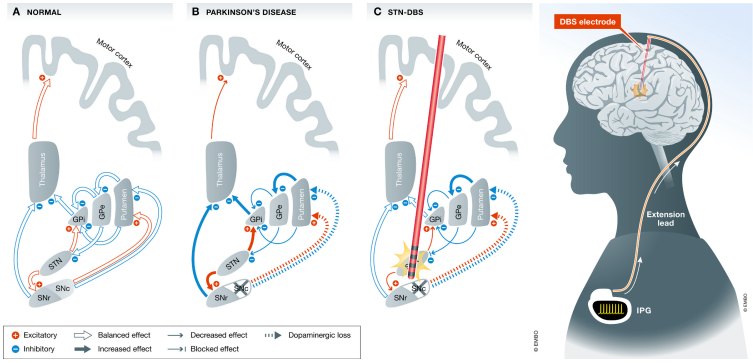


Figure 2: Motor pathways and subthalamic deep brain stimulation.
Left panel: Cortico-basal ganglia-thalamo-cortical loop. Visualization of electrophysiological signals controlling movement. Excitatory signals activate the receiving brain structure, inhibitory signals slow down the receiving brain structure. Right panel: DBS electrode located in basal ganglia region, connected via subcutaneous extension lead to the subcutaneous implanted IPG. DBS: deep brain stimulation, GPe: globus pallidus externa, GPi: globus pallidus interna, IPG: internal pulse generator, STN: subthalamic nucleus, SNc: substantia nigra pars compacta, SNr: substantia nigra pars reticulare.

Over the last decades, scientific advances, in for example genetics and cell biology, increased the fundamental understanding of PD immensely. And although this led to more and diverse potential treatment targets, it also revealed many new questions about a highly complex, multifactorial disease process. Generally, circa 90% of PD cases do not have a purely genetic or environmental cause, and is 'idiopathic' PD. Currently, it is assumed that a complex combination of genetic and

environmental factors lead to the intracellular accumulation of a protein, alpha-synuclein, and eventually to cell death[16,17]. This cell death typically occurs in dopamine-producing cells in the substantia nigra pars compacta (SNc), and results in a dopamine deficit which is the core of PD pathophysiology[18].

1.2 Parkinson symptomatology

The dopamine deficit causes a wide spectrum of symptoms in PD patients, consisting of motor and non-motor symptoms[19]. The work in this thesis focusses on motor symptoms, and thus non-motor symptoms will only be touched upon briefly. Further, it is important to be aware of the diversity and variation of symptomatology between PD patients[20], as this is important for individualized monitoring and treatment.

Bradykinesia is defined as slowness of movement, and hypokinesia is defined as decreased spontaneous movement, or reduced amplitude of movement [21]. It affects both arms and legs, and can manifest as full body bradykinesia, including decreased facial expression. Bradykinesia is the core symptom in diagnosing the parkinsonian syndrome according to the UK Parkinson's Disease Society Brain Bank clinical
diagnostic criteria (besides bradykinesia, rigidity, tremor or postural instability is required) [22]. It typically starts unilateral to become bilateral. Both the severity and the duration of bradykinesia worsen over the course of the disease, cooccurring with significant motor disabilities, and impaired QoL[23].
Rigidity is defined as stiffness of limb-movement. These symptoms often present simultaneously and are described as an akineto-rigid-syndrome.
Bradykinesia and rigidity respond well on early dopaminergic therapy, but over time complications such as motor fluctuations occur [24].

Tremor is one of the best-visible PD symptoms. It is defined as an involuntary shaking of the hand, arm, leg, foot, or neck. Typically, a Parkinsonian tremor is a resting tremor, starting when the patient is not using his hand or affected limb, and often increases in the presence of stress. This in contrast with an essential tremor, which occurs during voluntary movement. Tremor usually starts unilateral and gradually becomes bilateral, and both on upper and lower extremities.
In a subset of PD patients, tremor is the most-present motor symptom in a relative absence of other motor symptoms. This subtype of PD is described as tremor-dominant, and these patients often show slower disease progression and less functional-disability[18].

PD patients often suffer from *postural instability and gait problems*, typically in later disease stages. A typical form of gait disturbance is freezing of gait. This is defined as an inability to proceed or start walking. Also, abrupt ending of a walking episode can be disturbed in advanced PD[18].

A wide spectrum of *non-motor symptoms* characterizes PD, such as sleep disturbance and fatigue, pain, autonomic symptoms (constipation, urinal dysfunction, bowel incontinence, sexual difficulties), cognitive (memory and concentration difficulties), and neuropsychiatric disturbances (depression, anxiety, apathy) [19,25]. A variety of non-motor symptoms, such as limb pain, constipation, sleep disturbance, erectile dysfunction, and depression, typically precede the onset

of motor symptoms and PD diagnosis during a 'pro-dormal-phase', which can exist up to 20 years or more [18,25].
Since non-motor symptoms are less discriminative for PD, they are notoriously less well-known and neglected by clinicians, caregivers, and even patients [25]. Evidence that non-motor symptoms might influence patients' QoL more than motor symptoms is growing and underlines the importance of non-motor symptoms in PD[25,26]. This highlights the importance of adequate monitoring of non-motor symptoms in clinical practice to adjust therapies and symptom relief.

Motor fluctuations are a hallmark of PD progression and are characterized by daily fluctuations of therapeutic effectiveness on motor symptoms, mainly bradykinesia and rigidity (figure 3). They occur in half of the patients within the first decade after diagnosis [23,27]. After an initial period in which medication has a satisfactory effect, the motor symptoms become increasingly refractory to dopaminergic dosages. This process is called 'wearing-off'. Moments with satisfactory effect on motor symptoms are called ON-medication states, whereas moments with disabling motor symptoms are called OFF-medication states [24] (figure 3). With disease progression, higher dopamine-dosages are needed to realize sufficient ON-medication time, which often lead to levodopa induced dyskinesia. Higher levodopa dosages also correlate with the occurrence of motor fluctuations [28]. The theory that an early start of levodopa treatment leads to more motor fluctuations is dismissed [24].
Motor fluctuations have an important impact on a patient's QoL and are associated with worse ADL performance [29]. In addition to adequate monitoring of non-motor symptoms, implementing strategies to retrieve patterns in response-fluctuations will help the clinician and patient to adjust therapies and symptom relief.

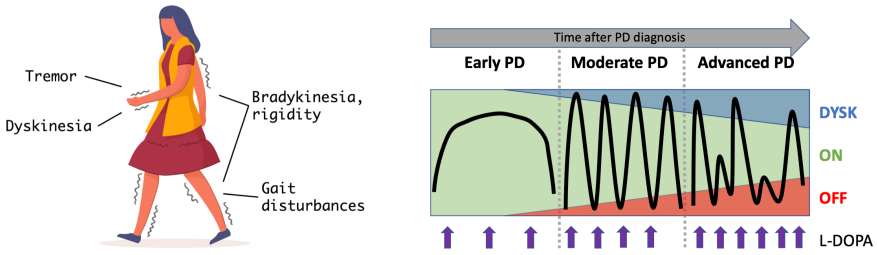


Figure 3: Motor symptoms and their fluctuations over time
Left panel: Overview of core motor symptoms in Parkinson's disease. Right panel: Fluctuating dopamine level in blood. The dopamine level (black line) fluctuates under influence of the intake of dopaminergic medication several times a day (L-DOPA, purple arrows). Over time, the time spent with good motor symptom control (ON-state) decreases. Parallel, the time spent with motor symptoms (such as bradykinesia and tremor) increases, due to a too low dopaminergic level (OFF-state). Also, the time spent with burdensome dyskinesia increases, due to a too high dopaminergic level (DYSK).

1.3 Treatment of motor symptoms

Although recent scientific advances in neuroprotective therapies using monoclonal antibodies give reason for 'cautious optimism', current PD treatment focusses on symptom-relief via dopamine-replacement rather than curation [17].

Dopaminergic medication

Dopamine-replacement therapy is centered around oral levodopa, a dopamine precursor introduced as PD therapy in the 1960s [17]. Levodopa crosses the blood-brain-barrier as a precursor and will be decarboxylated into dopamine, in order to stimulate neural dopamine receptors. The characteristic short pharmacokinetic half-life time of one to three hours [30], requires multiple daily intakes to maintain sufficient dopamine levels.

Current state-of-the-art policies advice to start dopaminergic treatment early, even when disabilities are still neglectable [24]. To increase levodopa efficacy, complementary drugs as dopamine agonists and monoamine-oxidase (MAO)-B inhibitors are given.

Over time patients need higher pharmacological dosages to yield satisfactory effects. High levodopa dosages are more likely to result in complications such as motor fluctuations (meaning OFF-medication periods), and levodopa induced dyskinesia. The latter is the occurrence of involuntary, uncontrolled movements and is believed to be the result of peak-levels of dopamine [24].

Deep Brain Stimulation history and current practice

Deep Brain Stimulation (DBS) therapy evolved from ablative surgery which was performed in the 1980s to structurally lesion the basal ganglia to mimic motor improvement observed after vascular basal ganglia lesions [31]. In 1987, Benabid et al introduced DBS as a reversible alternative for the irreversible structural lesioning [1]. Since then, the scientific and clinical DBS field emerged to several indications, and its working mechanism is hypothesized to rely on network-effects, rather than local effects [32,33].

Subthalamic nucleus (STN) and globus pallidus interna (GPi) DBS are safe, and effective therapies for severe Parkinsonian motor symptoms and dopaminergic motor complications, even after relatively short disease durations [2–5,7]. Cognitive impairment, depression, or neurosurgical contraindications can indicate intratestinal levodopa-gel infusion pump as an alternative [34]. PD patients considered for DBS will be multidisciplinary assessed by a neurologist, neurosurgeon, nurse, neuropsychologist, and neuropsychiatrist. Disease severity, levodopa efficacy, cognitive functionality, age, and social and private circumstances will determine the multidisciplinary decision-making.

5 to 10% of the PD population will be considered for DBS somewhen in their disease process. This leads annually to 125 to 300 expected DBS implantations in the Netherlands during the next decade [35].

DBS electrodes are surgically implanted in the STN or GPi with the aid of preoperative trajectory planning based on magneto-resonant images (MRI) and intraoperative stereotaxis (3-d navigation). Intraoperative electrophysiological recordings help to identify the target brain structure. Often patients stay awake during parts of the implantation to test whether stimulation results in symptom relief [36].

Efficacy of DBS for PD

STN-DBS is expected to decrease motor severity, assessed on MDS-UPDRS III, with one third in off-medication moments. Dopaminergic medication dosages are expected to halve, leading to less

dopaminergic induced dyskinesia. Time spent in on-medication state is expected to increase with several hours a day [2,3,5,6].
A subset of patients, circa 30%, still experiences a suboptimal motor response one-year after STN-DBS [3,6], and motor response appears to decrease on a mid-long term of 5 till 10 years [7]. The efficacy trade-off between motor response and adverse effects remains a challenge [8].

Approaches to improve DBS outcome in PD
Preoperative patient selection is a crucial part of DBS care. Numerous studies showed that preoperative more severe and longer-lasting motor symptoms, large motor response on dopaminergic medication, and an impaired QoL, correlate best with postoperative motor or QoL improvement [37–40]. These studies typically applied correlative statistics on group level.
The current computational possibilities, such as clinical data science including machine learning, enable clinical research to include more data and unravel more patterns than traditional statistics [9,41]. Also, machine learning delivers prediction models generating individual outcome probabilities usable in prospective settings.
Besides the abovementioned clinical variables [42], DBS outcome prediction in PD is explored based on motion sensing [43], structural imaging [44], connectivity imaging [45], preoperative neurophysiology [46], and genetic profiling [47].
Machine learning-driven individual DBS outcome prediction is therefore a promising tool which requires further exploration of its feasibility and validity. A realistic, useful, and impactful prediction tool for clinicians should always be the starting point of this research [11,48].

Postoperatively, DBS outcome can be improved by optimizing stimulation paradigms. Adaptive, or closed-loop, DBS (aDBS) is such an example and adjusts the amount of stimulation (aDBS output) based on a biomarker reflecting the patient's clinical state and thus needs (aDBS input). It is suggested to improve efficacy and efficiency of DBS in PD with less adverse effects and a longer battery life [49]. Besides tailoring the temporal character of DBS via aDBS, directional steering is suggested to adapt the loco-spatial character of DBS according to structures surrounding the implanted electrode causing adverse effects [50].

2. Real-life (naturalistic) Parkinson monitoring

Assessment of PD symptoms to adjust therapy and to provide optimal symptom relief is complicated by the heterogenous and fluctuating character of PD symptoms. The current gold standards forming the backbone of clinical PD evaluation are criticized for (among others) a lack of real-life representability. This, combined with emerging technological possibilities, led to great scientific and commercial interest in PD monitoring tools solutions [51–54]. Here, we will summarize the current state of PD monitoring and its limitations and focus on the advances made to overcome these limitations.

2.1 Traditional symptom assessments and their limitations

PD symptomatology is traditionally assessed with tools such as the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [55], the Hoehn and Yahr scale (HY), the Abnormal Involuntary Movement Score (AIMS), and non-motor scales. The Parkinson Disease Quality of life questionnaire (PDQ)-8 and the PDQ-39 specifically assess QoL in PD. The MDS-UPDRS is most used in clinical practice and as gold standard in therapy efficacy evaluations. It consists of four parts addressing cognitive and non-motor symptoms (part I), the ability to

perform activities of daily life (part II), motor symptoms (part III), and treatment induced side effect (part IV).
The MDS-UPDRS part III has to be conducted in-person by a trained-clinician who structurally interprets the assessed motor symptoms. This is a labor-intensive assessment and cannot be repetitively used over longer periods of time. The other MDS-UPDRS parts, the other tools, and the PDQ's are questionnaire-based, and cover symptoms over weeks or months. Self-report diaries to monitor motor states (OFF, ON, ON with dyskinesia) in PD patients' daily-life will be discussed later in more detail [56,57].
None of these tools facilitate the repetitive detection of short-term (non-)motor PD fluctuations in a real-life situation, over a longer period of time.

2.2 C*ommercial and scientific advances in PD monitoring*

Since the 1990s, motion sensors are investigated to monitor activity and symptomatology of PD patients [58]. Since then, common electronical devices able to capture real-life data have become increasingly presence. Currently, PD monitoring is of interest to many scientific and commercial investigators [51,59], including one of the largest tech-companies worldwide [53]. Specific guidelines from the Movement Disorders Society aim to guide these scientific efforts and advice non-proprietary algorithms which are validated under real-life conditions in PD patients, and devices without extra patient burden [52].
So far, devices are reported to be very successful in discriminating symptom severities in controlled settings, but the translation to symptom detection in uncontrolled, real-life settings is more challenging [51,60]. The first randomized controlled studies investigating the clinical impact of clinical decision-making augmented with real-life passive monitoring are currently being performed [61].
Descriptive and anecdotal reports about the first FDA-approved device for passive monitoring, the Personal KinetoGraph (PKG), are available. They describe promising augmentation of clinical decision making [62–64], and in particular improved detection of wearing-off effects [65]. In contrast, its validity over short time windows, such as one hour [66], and on the quality of current commercial validation processes are disputed [67].
In the meanwhile, scientific efforts focus on creating open-source and reproducible PD monitor methods [68,69]. Methodological details of motion sensor monitoring for PD will be discussed in paragraph 3.

2.3 Objective naturalistic PD monitoring: motion sensors

Accelerometers and gyroscopes are the two motion sensors applied in this thesis.
Accelerometers measure acceleration, velocity change (meters per second) per second, in the three axes X, Y, and Z (figure 4). Their functionality relies on three magnetic field components. Each of them produces a variable electronical signal under influence of acceleration in a specific axis.
Gyroscopes measure angular rotation, angular degrees per second, over the three axes roll, jaw, and pitch. They exist of 'tolling spin'-like components which rotate under influence of movement in their specific axis. The movement of the tolling-spin components is translated into degrees turned per second.

Signal processing of motion sensor data

Motion sensors result in high-frequency, e.g. 125 Hz, time-series data. These three signals, one signal per axis, have to be translated into meaningful data points for movement analysis. This is done via the extraction of features which describe relevant characteristics of movement.

The time window over which every feature is calculated is defined as the feature window. For example, a data set containing 15 minutes of tri-axial data (125 Hz) is used to extract 20 different features per axis, with a feature window of 1 second. The result is a data set of 60 features (20 features per axis) with 900 values per feature (15 minutes is 900 seconds, 1 value per second). Features can be extracted from the temporal domain, containing the acceleration or rotation (y-axis) over time (x-axis). Examples are features describing maxima, minima, variances, distributions, number of peaks, or smoothness. Features can also be extracted from the spectral domain, containing spectral power (y-axis) per frequency band (x-axis). This is calculated per time window (for example per second) and can be subsequently expressed as power recorded in the beta-frequency over time. A temporal domain signal is created through a Fast Fourier Transform (FFT) of the original temporal domain signal, which is a common signal processing method.

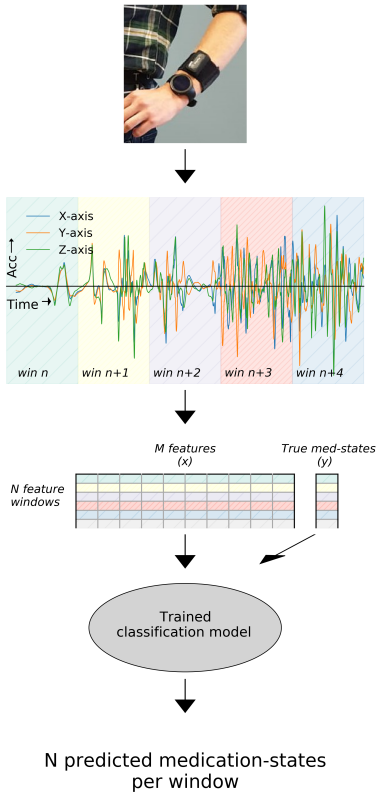


Figure 4: Workflow from motion sensor to medication state prediction.
The wrist-worn motion sensor (first panel) records a tri-axial accelerometer signal (can also be rotation) (second panel). Per feature window (color-shaded 'win's, in total here M), M features are extracted, and a true medication-state label is collected (third panel). A machine learning classifier is trained to recognize which medication-state label (y) is characterized by which patterns in accelerometer-features (x) (fourth panel). The model can be applied on new accelerometer data to predict to which medication-state the new feature windows most likely belong (fifth panel).

Translation of motion sensor data to Parkinson symptomatology

The goal of motion sensing in PD is to discriminate between different clinical symptom severities based on extracted features. Typically, PD patients perform experiments in which they perform activities which represent daily-life activities in different symptom states. If the extracted features will be discriminative enough, machine learning models can identify feature patterns corresponding to each symptom state. After learning these patterns, the machine learning models can recognize the patterns in motion sensor features which are not analyzed before. This way, the models can predict to which symptom state the new motion sensor data belong.

2.4 Subjective naturalistic PD monitoring: motor diaries and experience sampling method

Motor diaries still play an important role as gold standard for motor symptoms and fluctuations during naturalistic symptom monitoring in for example PD monitor device development. Classically, patients self-report their dopaminergic-medication states and presence of (burdensome) dyskinesia per 30 minutes in Hauser-diaries [56]. Electronical motor diaries (eDiaries) have been shown to be a reliable alternative to written diaries [70]. Experience sampling method (ESM), also known as ecological momentary assessment, can be applied as an eDiary to repetitively collect subjective patient experiences in natural environments [71]. Smartphone-based ESM presents an identical questionnaire to a patient, multiple times a day. Every time, the patient is asked to answer the question according to how they feel at that specific moment. To prevent bias through procrastination, every questionnaire can only be opened and completed within e.g. 15 minutes after notification. The fluctuations and trend in answers within days or between days provide information about the questioned symptom. ESM is validated and commonly used for several psychological or psychiatric indications [72,73]. Recently, eDiaries such smartphone-based ESM, are also introduced to give insights in naturalistic subjective experiences of PD patients [52,74]. Practical feasibility of ESM in a PD population has yet to be proven, and there is no ESM questionnaire described specifically for PD. Also, the use of PD eDiaries has not been reported in the motor monitoring literature so far.

3. Clinical data science methods

3.1 Terminology and definitions

Clinical data science is the term for the combined knowledge and skills regarding three domains according to the Drew Conway's Venn diagram [75]. First, knowledge of applied statistics and mathematics as commonly applied in scientific research is required. Second, knowledge of computer science and programming is required. Third, and last but not least, clinical expertise is required. Combined, these skills can reveal patterns between variables which cannot be disentangled without the computational power of computers [76]. Often, this boils down to a clinical question and hypothesis, involving predictors and outcomes of a specific process or treatment. Predictors are variables which serve as input for an analyzing model. Some examples are clinical variables such as demographic and disease specific characteristics, physiological time series such as heartrates, acceleration values, and neurophysiological assessments, or structural and functional neuroimaging. Both predictors and outcome variables can be continuous scores or categorical classes.

Clinical decision support systems (CDSS) are tools which provide a clinician with data-driven information to augment clinical decision making. This can be for example a probability of a diagnosis, a disease progression, or the effectiveness of a treatment. The provided information by a CDSS is often based on prediction models using machine learning algorithms. Often these models are not self-learning, and the algorithms are once developed and validated, to remain unchanged until a next development and training phase.

Supervised and unsupervised machine learning are two types of algorithms which differ in the way the model is trained. A *supervised machine learning* model is provided with the true labels of a *training data set*. The model investigates patterns between numerous predictors and the

corresponding outcome variable. Then, after the model learned the patterns between predictors and outcome, the model is provided with a new *test data* set, containing only predictors. The model will generate outcome predictions based on the learned patterns, and these outcome predictions are compared with the true outcome variables to evaluate the performance of the model. An *unsupervised machine learning* model is provided with predictors. The model compares all data points based on their predictors. Based on the differences and similarities, all data are clustered into different groups. Afterwards, additional analyses have to reveal how well these clusters align with the actual clinical outcome scores or classes. In this thesis, only supervised models are applied.

Classification models are machine learning models which use a categorical outcome score as outcome. Models analyzing continuous outcome scores are called regression models,but will not be applied and discussed in this thesis. In supervised classification models, there is a tradeoff between complexity and computational power on the one hand, and simplicity and interpretability on the other hand. In the era of big data and incremental powerful computers, many machine learning applications use complex models. However, there is a call for the use of simple, interpretable machine learning models in health care applications [77]. For many clinical prediction challenges, simple classification models are as good as complex models, and outperform them regarding interpretability [78]. The latter is an important factor in relying on machine learning, data-driven results with regard to clinical decision making which directly affects health care for patients.

Outcome probabilities generated by a classification model range from zero (0) to one (1). This probability value indicates how likely a specific data points belongs to the clinical outcome class which is defined as '1'. The eventual performance of the model is dependent on the threshold which is used to accept probabilities to be true (to belong to class 1). For example, when probabilities higher than 0.05 will be accepted, almost all data points will be predicted as '1'. Contrarily, whether probabilities higher than 0.90 will be accepted, most of the probabilities will be predicted as '0'. The optimal threshold will be dependent on the clinical utilization of the model and is strongly related to the clinical importance of false positive and false negative predictions. To evaluate the general performance of a model, regarding all possible probability thresholds, a so-called receiver operator characteristic (ROC) is used.

Predictive metrics are used to evaluate how well a classification model distinguishes between the outcome classes of data points based on their predictors. A general metric for model performance, without choosing an exact threshold of probability acceptance, is the *area under the receiver operator curve (AUC)*. The ROC visualizes for every probability threshold, the hypothetical true positive rate (TPR), and false positive rate (FPR). Therefore, a probability threshold is applied on all outcome probabilities, leading to predicted classes. The distribution of predicted classes versus actual classes are visualized in a confusion matrix, displaying *true negatives* (TN): predicted negative, true outcome negative; *false negatives* (FN): predicted negative, true outcome is positive; *true positives* (TP): predicted positive, true outcome positive; *false positives* (FP): predicted positive, true outcome negative. The TPR, y-axis of ROC, is defined as the part of true positives who was correctly predicted to be true; TP / (TP + FN). The FPR, x-axis of ROC, is defined as the part of true negatives who was falsely predicted to be true; FP / (FP + TN).

Several metrics exist to evaluate prediction models and their performance in specific clinical utilization, or with specific thresholds for probability acceptance. They all have their own pros and cons, and the clinical hypothesis of a model will define which metrics are most suitable. The *classification accuracy* is representing the amount of correct classified data points, compared with the total amount of data points, (TN + TP) / (TN+FN+TP+FP). The *positive predictive value (PPV)* is the part of positive predictions that is actual positive, TP / (TP + FP). The *negative predictive value (NPV)* is the part of negative predictions that is actual negative, TN / (TN + FN). The *sensitivity (or recall)* is the part of actual positives which was predicted correctly, TP / (TP + FN), and is equal to the TPR. The *specificity* is the part of negatives that was predicted correctly TN / (TN + FP), and is equal to (1 – FPR).

3.2 Challenges to translate AI into clinical impact

For circa 30 years, scientists try to revolutionize clinical care with clinical data science, AI and machine learning [79]. Experience has proven that AI-applications are notoriously difficult to implement in clinical practice. Important reasons are the lack of mutual understanding and collaboration between clinicians and computer scientists [48]. Close collaborations and scientific reporting via dedicated guidelines should contribute to overcome this limitation [11,80].
Both Pencina and colleagues, and Kelly and colleagues outlined key considerations to develop impactful clinical prediction models from hypothesis up to clinical implementation [81]. These studies provide understandable statistical and conceptual insights and form an excellent bridge between the clinical and the computational perspectives. Regarding the choice of mathematics, there is a strong call for the use of interpretable, machine learning models for clinical applications [77,78,82]. In this discussion, the analogy black-box versus white-box refers respectively to models without insight in the establishment of predictions versus models which give insight in predictor importance for example.

The work in this thesis aims to implement clinical data science methodologies for preoperative DBS outcome prediction and motion sensor monitoring, without losing sight of the clinical relevance and the mutual understanding of clinicians and data scientists.

4. Problem statements and outline of this thesis

PD care can be improved if (wearable or clinical) patient data and the computational possibilities are understood and implemented correctly, and combined in a clinically meaningful manner. In this thesis we will focus on the improvement of DBS care on two topics. First, not all patients experience a satisfying motor response postoperatively and, despite all efforts, we cannot identify these patients preoperatively [3,6]. Second, not all patients show a constant satisfying balance between beneficial motor response and adverse effects, and we cannot track this motor response in daily life [7,8,49].
Part A addresses the first challenge by predicting individual postoperative motor response, during the preoperative phase based on clinical variables. **Chapter 2** describes the development of DBS-PREDICT, a machine learning classification model to differentiate between weak and strong motor responders of subthalamic DBS. **Chapter 3** describes a multicenter validation of DBS-PREDICT, which is the first external validation of a prediction tool for individual DBS outcome.

Part B addresses the second challenge by exploring how continuous daily-life PD monitoring can contribute to (adaptive) DBS outcome monitoring. **Chapter 4** reviews the advances so far regarding aDBS for PD. **Chapters 5, 6, and 7** describe the development, feasibility, and proof-of-concept of electronic motor diaries as a tool to improve motion-sensor monitoring in PD. **Chapter 8** explores the feasibility of motion-sensor PD monitoring on short time windows, which is a requirement to evaluate the motor response to (adaptive) DBS in real-time rather than per day.
Chapter 9 provides a general discussion of all chapters, and **chapter 10** analyses the valorization and impact on future research and clinical practice of the work presented in this thesis. **Chapter 11 and 12** provide a concluding summary and valorization paragraph.

References

1. Benabid, A.-L., Pollak, P., Louveau, A., Henry, S. & De Rougemont, J. Combined (thalamotomy and stimulation) stereotactic surgery of the VIM thalamic nucleus for bilateral Parkinson disease. Stereotact. Funct. Neurosurg. 50, 344–346 (1987).
2. Deuschl, G. et al. A randomized trial of deep-brain stimulation for Parkinson's disease. N Engl J Med 355, 896–908 (2006).
3. Williams, A. et al. Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. Lancet Neurol 9, 581–91 (2010).
4. Schuepbach, W. M. et al. Neurostimulation for Parkinson's disease with early motor complications. N Engl J Med 368, 610–22 (2013).
5. Vitek, J. L. et al. Subthalamic nucleus deep brain stimulation with a multiple independent constant current-controlled device in Parkinson's disease (INTREPID): a multicentre, double-blind, randomised, sham-controlled study. Lancet Neurol 19, 491–501 (2020).
6. Weaver, F. M. et al. Bilateral deep brain stimulation vs best medical therapy for patients with advanced Parkinson disease: a randomized controlled trial. Jama 301, 63–73 (2009).
7. Janssen, M. L. et al. Subthalamic nucleus high-frequency stimulation for advanced Parkinson's disease: motor and neuropsychological outcome after 10 years. Ster. Funct Neurosurg 92, 381–7 (2014).
8. Little, S. & Brown, P. What brain signals are suitable for feedback control of deep brain stimulation in Parkinson's disease? Ann N Acad Sci 1265, 9–24 (2012).
9. Pedersen, M. et al. Artificial intelligence for clinical decision support in neurology. Brain Commun. 2, (2020).
10. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. Lancet 393, 1577–1579 (2019).
11. Mateen, B. A., Liley, J., Denniston, A. K., Holmes, C. C. & Vollmer, S. J. Improving the quality of machine learning in health applications and clinical research. Nat. Mach. Intell. 2, 554–556 (2020).
12. Parkinson, J. An Essay on the Shaking Palsy. (1817).
13. Dorsey, E. R., Sherer, T., Okun, M. S. & Bloem, B. R. The Emerging Evidence of the Parkinson Pandemic. J Park. Dis 8, S3-s8 (2018).
14. Leisman, G. & Melillo, R. The basal ganglia: motor and cognitive relationships in a clinical neurobehavioral context. Rev. Neurosci. 24, 9–25 (2013).
15. Brunenberg, E. J. L. et al. Structural and Resting State Functional Connectivity of the Subthalamic Nucleus: Identification of Motor STN Parts and the Hyperdirect Pathway. PLOS ONE 7, e39061 (2012).
16. Michel, P. P., Hirsch, E. C. & Hunot, S. Understanding Dopaminergic Cell Death Pathways in Parkinson Disease. Neuron 90, 675–91 (2016).
17. Jankovic, J. & Tan, E. K. Parkinson's disease: etiopathogenesis and treatment. J Neurol Neurosurg Psychiatry 91, 795–808 (2020).
18. Kalia, L. V. & Lang, A. E. Parkinson's disease. Lancet 386, 896–912 (2015).
19. Marras, C. & Chaudhuri, K. R. Nonmotor features of Parkinson's disease subtypes. Mov Disord 31, 1095–102 (2016).
20. Titova, N., Padmakumar, C., Lewis, S. J. G. & Chaudhuri, K. R. Parkinson's: a syndrome rather than a disease? J Neural Transm Vienna 124, 907–914 (2017).
21. Bologna, M., Paparella, G., Fasano, A., Hallett, M. & Berardelli, A. Evolving concepts on bradykinesia. Brain 143, 727–750 (2020).
22. Gibb, W. R. & Lees, A. J. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry 51, 745–52 (1988).
23. Fasano, A. et al. Characterizing advanced Parkinson's disease: OBSERVE-PD observational study results of 2615 patients. BMC Neurol 19, 50 (2019).
24. de Bie, R. M. A., Clarke, C. E., Espay, A. J., Fox, S. H. & Lang, A. E. Initiation of pharmacological therapy in Parkinson's disease: when, why, and how. Lancet Neurol. 19, 452–461 (2020).
25. Todorova, A., Jenner, P. & Ray Chaudhuri, K. Non-motor Parkinson's: integral to motor Parkinson's, yet often neglected. Pract. Neurol. 14, 310 (2014).
26. Kuhlman, G. D., Flanigan, J. L., Sperling, S. A. & Barrett, M. J. Predictors of health-related quality of life in Parkinson's disease. Parkinsonism Relat. Disord. 65, 86–90 (2019).
27. Kim, H.-J. et al. Motor Complications in Parkinson's Disease: 13-Year Follow-up of the CamPaIGN Cohort. Mov. Disord. 35, 185–190 (2020).
28. Bjornestad, A. et al. Risk and course of motor complications in a population-based incident Parkinson's disease cohort. Park. Relat Disord 22, 48–53 (2016).
29. Hechtner, M. C. et al. Quality of life in Parkinson's disease patients with motor fluctuations and dyskinesias in five European countries. Parkinsonism Relat. Disord. 20, 969–974 (2014).

30. Information, N. C. for B. PubChem Compound Summary for CID 6047, Levodopa. https://pubchem.ncbi.nlm.nih.gov/compound/Levodopa (2021).
31. Benabid, A. L. et al. Long-term electrical inhibition of deep brain targets in movement disorders. Mov Disord 13 Suppl 3, 119–25 (1998).
32. Jakobs, M., Fomenko, A., Lozano, A. M. & Kiening, K. L. Cellular, molecular, and clinical mechanisms of action of deep brain stimulation—a systematic review on established indications and outlook on future developments. EMBO Mol. Med. 11, e9575 (2019).
33. Ashkan, K., Rogers, P., Bergman, H. & Ughratdar, I. Insights into the mechanisms of deep brain stimulation. Nat. Rev. Neurol. 13, 548–554 (2017).
34. Odin, P. et al. Collective physician perspectives on non-oral medication approaches for the management of clinically relevant unresolved issues in Parkinson's disease: Consensus from an international survey and discussion program. Park. Relat Disord 21, 1133–44 (2015).
35. Stimulation, E. D. B. Behoefteraming Deep Brain Stimulation (versie 3). (2019).
36. Temel, Y. et al. Single electrode and multiple electrode guided electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease. Neurosurgery 61, 346–55; discussion 355-7 (2007).
37. Kleiner-Fisman, G. et al. Subthalamic nucleus deep brain stimulation: summary and meta-analysis of outcomes. Mov Disord 21 Suppl 14, S290-304 (2006).
38. Daniels, C. et al. Is improvement in the quality of life after subthalamic nucleus stimulation in Parkinson's disease predictable? Mov Disord 26, 2516–21 (2011).
39. Schuepbach, W. M. M. et al. Quality of life predicts outcome of deep brain stimulation in early Parkinson disease. Neurology (2019) doi:10.1212/wnl.0000000000007037.
40. Cavallieri, F. et al. Predictors of Long-Term Outcome of Subthalamic Stimulation in Parkinson Disease. Ann. Neurol. doi:10.1002/ana.25994.
41. Beam, A. L. & Kohane, I. S. Big Data and Machine Learning in Health Care. JAMA 319, 1317–1318 (2018).
42. Frizon, L. A. et al. Quality of Life Improvement Following Deep Brain Parkinson's Disease: Development of a Prognostic Model. Neurosurgery (2018) doi:10.1093/neuros/nyy287.
43. Pulliam, C. L. et al. Motion sensor strategies for automated optimization of deep brain stimulation in Parkinson's disease. Park. Relat Disord 21, 378–82 (2015).
44. Younce, J. R., Campbell, M. C., Perlmutter, J. S. & Norris, S. A. Thalamic and ventricular volumes predict motor response to deep brain stimulation for Parkinson's disease. Park. Relat Disord (2018) doi:10.1016/j.parkreldis.2018.11.026.
45. Shang, R., He, L., Ma, X., Ma, Y. & Li, X. Connectome-Based Model Predicts Deep Brain Stimulation Outcome in Parkinson's Disease. Front. Comput. Neurosci. 14, (2020).
46. Victor J. Geraedts. Right on Track: Towards improving DBS patient selection and care. (Leiden University, 2020).
47. Rizzone, M. G., Martone, T., Balestrino, R. & Lopiano, L. Genetic background and outcome of Deep Brain Stimulation in Parkinson's disease. Parkinsonism Relat. Disord. 64, 8–19 (2019).
48. Higgins, D. & Madai, V. I. From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. Adv. Intell. Syst. 2, 2000052 (2020).
49. Little, S. et al. Adaptive deep brain stimulation in advanced Parkinson disease. Ann Neurol 74, 449–57 (2013).
50. Schnitzler, A. ; B., M. ;. Verhagen, L. ;. Groppa, S. ;. Cheeran, B. ;. Karst, E. ;. Defresne, F. ;. Vesper, J. Directional versus omnidirectional Deep Brain Stimulation for Parkinson's disease: Results of a prospective, blinded, multi-center, single-arm crossover study [abstract]. in (2019).
51. Thorp, J. E., Adamczyk, P. G., Ploeg, H. L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. Front Neurol 9, 1036 (2018).
52. Espay, A. J. et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 34, 657–663 (2019).
53. Powers, R. et al. Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease. Sci. Transl. Med. 13, eabd7865 (2021).
54. Horne, M. K., McGregor, S. & Bergquist, F. An objective fluctuation score for Parkinson's disease. PLoS One 10, e0124522 (2015).
55. Goetz, C. G. et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 23, 2129–70 (2008).
56. Hauser, R. A., McDermott, M. P. & Messing, S. Factors associated with the development of motor fluctuations and dyskinesias in Parkinson disease. Arch. Neurol. 63, 1756–1760 (2006).
57. Jansen, E. N. H., Meerwaldt, J. D., Heersema, T., Manen, J. V. & Speelman, J. D. Open Multicenter Trial with Madopar HBS in Parkinsonian Patients. Eur. Neurol. 27, 88–92 (1987).
58. van Hilten, J. J., Middelkoop, H. A., Kerkhof, G. A. & Roos, R. A. A new approach in the assessment of motor activity in Parkinson's disease. J Neurol Neurosurg Psychiatry 54, 976–9 (1991).
59. Warmerdam, E. et al. Long-term unsupervised mobility assessment in movement disorders. Lancet Neurol. 19, 462–470 (2020).

60. Galperin, I. et al. Associations between daily-living physical activity and laboratory-based assessments of motor severity in patients with falls and Parkinson's disease. Park. Relat Disord 62, 85–90 (2019).
61. Rodriguez-Molinero, A. Monitoring of Mobility of Parkinson's Patients for Therapeutic Purposes - Clinical Trial (MoMoPa-EC). NCT04176302 (2019).
62. Pahwa, R. et al. Role of the Personal KinetiGraph in the routine clinical assessment of Parkinson's disease: recommendations from an expert panel. Expert Rev Neurother 18, 669–680 (2018).
63. Joshi, R. et al. PKG Movement Recording System Use Shows Promise in Routine Clinical Care of Patients With Parkinson's Disease. Front Neurol 10, 1027 (2019).
64. Santiago, A. et al. Qualitative Evaluation of the Personal KinetiGraph TM Movement Recording System in a Parkinson's Clinic. J. Park. Dis. 9, 207–219 (2019).
65. Farzanehfar, P., Woodrow, H. & Horne, M. Assessment of Wearing Off in Parkinson's disease using objective measurement. J Neurol (2020) doi:10.1007/s00415-020-10222-w.
66. Ossig, C. et al. Correlation of Quantitative Motor State Assessment Using a Kinetograph and Patient Diaries in Advanced PD: Data from an Observational Study. PLoS One 11, e0161559 (2016).
67. Fasano, A. & Mancini, M. Wearable-based mobility monitoring: the long road ahead. Lancet Neurol. 19, 378–379 (2020).
68. Mahadevan, N. et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. NPJ Digit Med 3, 5 (2020).
69. Rochester, L. et al. A Roadmap to Inform Development, Validation and Approval of Digital Mobility Outcomes: The Mobilise-D Approach. Digit. Biomark. 4, 13–27 (2020).
70. Terroba-Chambi, C. et al. Open-Access Electronic Diary for Motor Fluctuation and Dyskinesia Evaluation in Parkinson Disease: Comparison With Paper Diary. Clin. Neuropharmacol. 41, 20–22 (2018).
71. Palmier-Claus, J. E. et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 123, 12–20 (2011).
72. Simons, C. J. et al. Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. Eur Psychiatry 30, 900–6 (2015).
73. Bailon, C. et al. Smartphone-Based Platform for Affect Monitoring through Flexibly Managed Experience Sampling Methods. Sens. Basel 19, (2019).
74. Broen, M. P. et al. Unraveling the Relationship between Motor Symptoms, Affective States and Contextual Factors in Parkinson's Disease: A Feasibility Study of the Experience Sampling Method. PLoS One 11, e0151195 (2016).
75. Kubben, P. L. Introducing "computational neurosurgery". Surg. Neurol. Int. 8, (2017).
76. Kubben, P., Dumontier, M. ,. Dekker, A. Fundamentals of Clinical Data Science. (Springer International Publishing, 2018).
77. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206 (2019).
78. Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110, 12–22 (2019).
79. Shortliffe, E. H. The adolescence of AI in Medicine: Will the field come of age in the '90s? Artif. Intell. Med. State---Art Future Prospects 5, 93–106 (1993).
80. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17, 195 (2019).
81. Pencina, M. J., Goldstein, B. A. & D'Agostino, R. B. Prediction Models - Development, Evaluation, and Clinical Application. N Engl J Med 382, 1583–1586 (2020).
82. Shah, N. H., Milstein, A. & Bagley Ph, D. S. Making Machine Learning Models Clinically Useful. Jama (2019) doi:10.1001/jama.2019.10306.

Part A:

Prediction of DBS motor response in Parkinson's disease

Chapter 2

Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease
-
Proof of principle in a retrospective cohort

Habets JGV, Janssen MLF, Duits AA, Sijben LC, Mulders AEP, De Greef B, Temel Y, Kuijf ML, Kubben PL, Herff C

Published in PeerJ 2020, 8:e10317
DOI: https://doi.org/10.7717/peerj.10317

Abstract

Introduction: Despite careful patient selection for subthalamic nucleus deep brain stimulation (STN DBS), some Parkinson's disease patients show limited improvement of motor disability. Innovative predictive analyzing methods hold potential to develop a tool for clinicians that reliably predicts individual postoperative motor response, by only regarding clinical preoperative variables. The main aim of preoperative prediction would be to improve preoperative patient counselling, expectation management, and postoperative patient satisfaction.

Methods: We developed a machine learning logistic regression prediction model which generates probabilities for experiencing weak motor response one year after surgery. The model analyses preoperative variables and is trained on 89 patients using a five-fold cross-validation. Imaging and neurophysiology data are left out intentionally to ensure usability in the preoperative clinical practice.
Weak responders (n = 30) were defined as patients who fail to show clinically relevant improvement on Unified Parkinson Disease Rating Scale II, III or IV.

Results: The model predicts weak responders with an average area under the curve of the receiver operating characteristic of 0.79 (standard deviation: 0.08), a true positive rate of 0.80 and a false positive rate of 0.24, and a diagnostic accuracy of 78%. The reported influences of individual preoperative variables are useful for clinical interpretation of the model, but cannot been interpreted separately regardless of the other variables in the model.

Conclusion: The model's diagnostic accuracy confirms the utility of machine learning based motor response prediction based on clinical preoperative variables.
After reproduction and validation in a larger and prospective cohort, this prediction model holds potential to support clinicians during preoperative patient counseling.

Introduction

Subthalamic nucleus deep brain stimulation (STN DBS) is a widely accepted therapy for Parkinson's disease (PD) patients in which dopaminergic replacement therapy is unsatisfactory.[1-4] In the majority of these patients, DBS can reduce motor symptoms or their fluctuations and thereby improve quality of life.[5] Despite careful patient selection, some patients still show limited or no improvement of motor fluctuations and quality of life.[5] Since the introduction of STN DBS, clinicians aimed to determine reliable predictors.[6]

Preoperative levodopa responsiveness of motor symptoms, severity of motor symptoms, and younger age are repeatedly reported as positive predictive factors for postoperative (Movement Disorders Society –) Unified Parkinson's Disease Rating Scale ((MDS-)UPDRS) motor improvement.[7] Contrarily, preoperative levodopa responsiveness is also reported to not predict STN DBS outcome.[8,9] Preoperative severe quality of life (QoL) impairment, more time spent in off-condition of dopaminergic medication, levodopa responsiveness, and low BMI are reported as positive predictive factors on postoperative QoL.[8,10-12] Reports on the predictive value of disease duration, daily levodopa dosage, postural and gait impairment, and non-motor symptoms all show conflicting results.[7,11,13,14] Comparison of reported motor outcome is hampered due to variance in assessment scales and assessments during varying dopaminergic states.[15]

These non-conclusive results maintain the need for a simple tool which neurologists can use in clinical practice to predict motor outcome after STN DBS for individual patients. To realize a usable and representative tool for the preoperative setting, our approach is limited to preoperative clinical variables. Preoperative prediction will always lack surgical information such as lead placement. This lack of information is inherent to any approach that aims to contribute to a better preoperative counselling.

Machine learning methods are increasingly used in medical practice to unravel patterns to improve understanding of clinical data.[16] Predictive machine learning models can be distinguished from traditional statistics by generating outcome predictions for new, individual patients, instead of correlations between pre- and postoperative variables on a group level. To ensure practical usability, clinical relevance, and interpretable results, the development and implementation of these models requires statistical, programming, and clinical expertise.[17] To add value to PD care, predictive analysis should improve challenging clinical decision making instead of reproduce valid clinical decisions.[18,19] Here, we report the development and proof-of-concept of a prediction model that generates probabilities for weak and strong motor response one year after STN DBS for individual PD patients based on preoperative clinical variables.

Materials and Methods

Study population

We considered patients who underwent STN DBS for PD in our academic neurosurgical centre between 2004 and March 2018. The surgical procedure is described in the Supplemental Material. We included 127 patients who completed one-year postoperative follow up during this period. We excluded patients who had missing UPDRS-III scores in their preoperative on-medication condition, or postoperative on-medication, on-stimulation condition.

The Medical Ethical Committee of Maastricht UMC+ approved this study (2018-0739). Informed consent was not obtained since the retrospective data was collected coded.

Pre- and postoperative variables

All available preoperative demographic data, disease specific data (disease onset, disease duration, levodopa equivalent daily dosage (LEDD)),[20] clinical performance scores ((MDS)-UPDRS, and Hoehn & Yahr (H&Y) scores), as well as relevant neuropsychological scores assessing executive functioning, in particular verbal fluency (semantic and lexical) and response inhibition (based on the interference score of the Stroop Colour Word Test) were incorporated. We left out imaging and neurophysiology data, to ensure the user-friendliness and accessibility in clinical practice during preoperative counselling. No analyses are required which ask software, hardware, or analysing knowledge.

All included preoperative clinical and neuropsychological scores were assessed in the on-medication condition and the available (MDS-)UPDRS III and H&Y scores in the off-medication condition were also included. Preoperative motor levodopa-responsiveness was calculated by subtracting UPDRS III scores in the off-medication condition with UPDRS III scores in the on-medication condition. Postoperative collected variables consist of UPDRS I, II, III and IV and H&Y scores in on-medication and on-stimulation conditions, UPDRS III in on-stimulation and off-medication conditions, and the performance on the verbal fluency and Stroop tests in on-stimulation and on-medication conditions. Both MDS-UPDRS and UPDRS scores were collected due to the variation in surgery dates among the population. To create uniform UPDRS scores, all MDS-UPDRS scores were recalculated to UPDRS scores.[15] Pre- and postoperative differences for UPDRS scores I until IV, H&Y scores, LEDD, and neuropsychological scores were calculated. Furthermore, we registered applied DBS voltage, frequency, and pulse width at one-year follow up. To compare DBS-settings, we computed the mean total electrical energy delivered (TEED).[21]

Prediction model

The machine learning prediction model uses multivariate logistic regression analyses. This logistic regression model distinguishes itself from (univariate) correlative regression models by generating outcome probabilities for individual patients (fig. 1). We focused on motor response as outcome and differentiated between 'strong responders' and 'weak responders'. To capture a wide spectrum of motor responders, improvement on UPDRS II, III and IV was evaluated. A strong responder is defined as a patient who showed a minimal clinically important difference (MCID) on UPDRS II, III, or IV in on-medication and on-stimulation condition one-year postoperative vs. preoperative on-medication condition (see fig. 2). MCID was defined as more than 3, 5, and 3 points improvement for UPDRS II, III and IV respectively, based on a literature review (see Supplemental Material). Patients who improved more than the MCID on UPDRS II or IV, but showed a deterioration on UPDRS III of more than the MCID and the yearly natural disease progression together 7 points, were defined as weak responders [22,23].

The prediction model uses the following available preoperative variables to generate an outcome probability: gender, age at DBS, PD duration at DBS, age at PD onset, UPDRS I, II, III and IV in on-medication condition, motor levodopa response, H&Y scale in on- and off-condition, the Stroop interference score, the verbal fluency scores, and the LEDD.

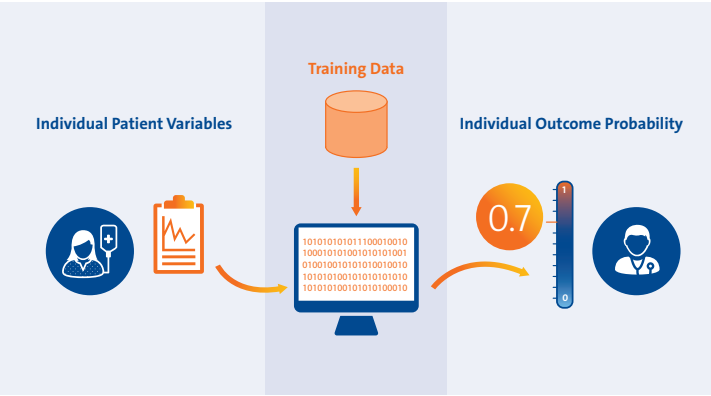


Figure 1: Overview of prediction approach.
Workflow of the prediction model as a preoperative counselling tool. The preoperative individual patient variables are inserted in the prediction model, which is trained on the retrospective database ('Training Data'). The model calculates the probability to become a weak responder between 0 and 1, in this example 0.7. The clinician can use this probability to inform the patient during preoperative counselling.

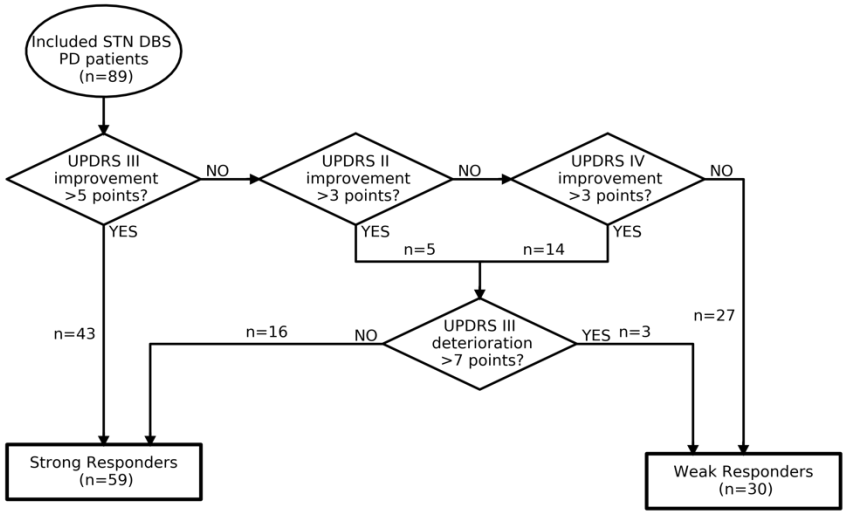


Figure 2: Decision flowchart of outcome categorization.
DBS, deep brain stimulation; PD, Parkinson's disease; STN, subthalamic; UPDRS, Unified Parkinson Disease Rating Scale.

The logistic regression model was fitted, i.e. trained, on the relation between preoperative variables and postoperative outcome categorization.[24] We evaluated the trained model with a 5-fold cross-validation. This cross-validation fits, i.e. trains, the model on 80% of the patients, the 'training data'. During this 'training phase', a weight, 'β', is assigned to every single preoperative variable, 'x'. The fitted model was then evaluated, i.e. tested, on the remaining 20% of patients in the database, the 'test data'. During this 'test phase', the preoperative variables of every individual patient in the test data were inserted in the model separately. The model generates an

outcome probability to become a weak responder for every individual patient. This probability was generated by a calculation of all 'x' values of the inserted patient with the corresponding weights (β) using the logistic function 1 / (1 + exp(- β * x)). The generated probabilities from the test data are compared with the actual outcome to test predictive accuracy. The 5-fold cross-validation repeats these phases 5 times until every patient was used for testing exactly once. The cross-validation leads to less limitations in sample size regarding number of considered predictive variables.[25] Still, the small number of patients on which the trained model is tested during every iteration in this 5-fold cross-validation is a limitation of this approach. Evaluating the average performance over the 5 iterations gives the best assumption of the predictive performance of the model. We chose logistic regression as a prediction model instead of a deep learning-based model due to the relatively small database size and the fact that the weight, or influence, of every preoperative variable can be interpret easily. This interpretation helps to generate an intuition what the prediction is based on.[26]

To use a certain prediction model in clinical practice, a threshold should be chosen to accept a probability. This means every probability above the threshold is regarded to be true (weak response in this model), and every probability below the threshold is regarded to be false (strong response in this model). The accuracy of the model is strongly dependent on the threshold. A common way to evaluate the overall performance of a prediction model is to plot the receiver operating characteristic (ROC). The ROC visualizes for different thresholds between 0 and 1 the corresponding true positive and false positive rates (fig 3A). Performance of prediction models is often expressed as the area under the curve (AUC) of the ROC (fig. 3A). In clinical practice, a threshold should be selected before the model can be used as a prospective application.

To understand which variables are important in the prediction model, we can explore the importance of every separate preoperative variable. Variable importance is expressed as 'weights'. To make these weights interpretable, they are converted to Odds Ratios by calculating exp(β), and normalized afterwards. These normalized Odds Ratios are called 'relative influences', and they denote the change in probability to be a weak responder when the respective variable increases 1 unit, and all other variables stay equal (fig 3A).

Comparative descriptive analysis between preoperative and postoperative variables and between weak and strong responders are performed with Mann-Whitney-U-tests.
To facilitate prediction models, we imputed missing data-points in preoperative variables. (For further explanation on the Random Forest imputations applied on preoperative variables, please see Supplementary Material). To prevent imputations of variables that are the target of prediction, we did not impute postoperative variables. Analysis is performed in Python Jupyter Notebook 3 (Jupyter Team, https://jupyter.org, revision fe7c2909) using packages pandas (version 1.0.4), Numpy (version 1.16.4), scikit-learn (version 0.21.2), and Scipy (version 1.3.0). We report our findings according to the TRIPOD Checklist for Prediction Model Development.[27]

Results

Preoperative and postoperative variables

We included 89 patients with a well-documented one year follow up after STN DBS, 37 patients were excluded due to missing data points in UPDRS III score in preoperative on-medication condition, or postoperative on-medication and on-stimulation condition. We report descriptive

statistics containing the original data (no imputed preoperative data). The total group showed statistically significant postoperative improvements in UPDRS III scores, both compared with preoperative on- and off-medication conditions, and in UPDRS IV scores. We observed a significant decrease in LEDD (table 1). Further, there was a significant deterioration in neuropsychological scores on a group level.

		Baseline characteristics [a]
Female sex (n)		37 (42%)
Age (years)		61 (8)
Disease duration (years)		10.7 (5.1)
Preoperative UPDRS III levodopa response		−18.6 (13.1)
Preoperative UPDRS III % levodopa response		−45.0 (38.0)
	Preoperative [a]	**1 year follow up** [a]
UPDRS I[b]	1.3 (1.3)	3.0 (3.1)
UPDRS II[b]	9.8 (6.6)	9.6 (5.5)
UPDRS III[b]	21.9 (12.5)	16.4 (9.9)[e]
UPDRS III[c]	39.1 (13.1)	16.4 (9.9)[e]
UPDRS IV[b]	5.5 (4.0)	2.8 (2.4)[e]
H&Y 1[b]	2 (2%)[d]	4 (3%)
H&Y 1.5[b]	2 (2%)	1 (1%)
H&Y 2[b]	13 (15%)	21 (30%)
H&Y 2.5[b]	34 (40%)	24 (34%)
H&Y 3[b]	25 (29%)	19 (27%)
H&Y 4[b]	9 (11%)	2 (3%)
H&Y 5[b]	–	–
Fluency total categories[b]	39.7 (9.4)	33.6 (9.8)[e]
Fluency total letters[b]	35.5 (10.8)	33.6 (11.9)[e]
Stroop interference[b]	56.1 (35.1)	76.7 (63.1)[e]
LEDD (milligrams)	1187 (619)	656 (510)[e]
TEED	–	134 (130)

Table 1: Preoperative and postoperative variables of total population.
H&Y: Hoehn & Yahr scale; LEDD: levodopa equivalent daily dosage; off-/on-med: off-/on-medication; off-/on-stim: off-/on-stimulation; TEED: total electrical energy delivered; UPDRS**:** Unified Parkinson Disease Rating Scale. a: Values are given as mean and standard deviation of the mean. b: Preoperative: on-medication, postoperative: on-medication and on-stimulation. c: Preoperative: off-medication, postoperative: off-medication and on-stimulation. d: Percentage of Hoehn and Yahr scales are relative based on the number of available data (pre: *n* = 85, post: *n* = 71). e: Significant difference with p-value < 0.05, calculated with Mann Whitney-U test

59 out of 89 patients were categorized as strong responders, 30 patients were categorized as weak responders (fig. 2). Postoperative clinical records until one-year follow-up were evaluated and surgical factors explaining weak response were ruled out for all weak responders. The groups had significant differences on all postoperative UPDRS scores and differences, except for the UPDRS III during on-stimulation and off-medication state (see table 2). We observed no significant or relevant differences between the groups regarding neuropsychological scores, LEDD, or TEED.

1 year follow up variables	**Strong responders, *n* = 59**[a]	**Weak responders, *n* = 30**[a]
UPDRS I[b]	2.2 (2.2)	4.5 (4.0) [f]
UPDRS I change[c]	0.4 (1.9)	1.5 (1.7) [f]
UPDRS II[b]	8.4 (4.7)	12.3 (6.1) [f]

UPDRS II change[c]	−2.8 (6.3)	5.6 (5.9) [f]
UPDRS III, on-med[b]	13.9 (7.5)	21.5 (11.9) [f]
UPDRS III change[c]	−11.9 (11.6)	7.3 (8.5) [f]
UPDRS III, off-med[d]	20.9 (13.1)	25.7 (6.7)
UPDRS III change[e]	−29.0 (13.8)	−12.4 (14.4) [f]
UPDRS IV[b]	2.4 (2.2)	3.5 (2.6) [f]
UPDRS IV change[c]	−4.1 (3.6)	0.1 (4.2) [f]
Fluency total categories[b]	33.8 (10.4)	33.4 (8.5)
Fluency total letters[b]	31.2 (11.5)	31.8 (12.8)
Stroop interference[b]	75.6 (66.4)	79.0 (55.7)
LEDD	622 (511)	717 (501)
LEDD change	−509 (472)	−577 (529)
LEDD change (%)	−40.7 (37.5)	−42.2 (35.2)
TEED	145 (151)	112 (69)

Table 2: Comparison of postoperative variables in groups with strong responders and weak responders.
LEDD: levodopa equivalent daily dosage; off-/on-med: off-/on-medication; off-/on-stim: off-/on-stimulation; TEED: total electrical energy delivered; UPDRS: Unified Parkinson Disease Rating Scale.
a: Mean (standard deviation). b: On-stimulation, on-medication at one-year follow up. c: Difference between on-medication and on-stimulation vs. preoperative on-medication. d: On-stimulation and off-medication at one-year follow up. e: Difference between on-stimulation and off-medication vs. preoperative off-medication. f: Significant difference with p-value < 0.05, calculated with Mann Whitney-U test

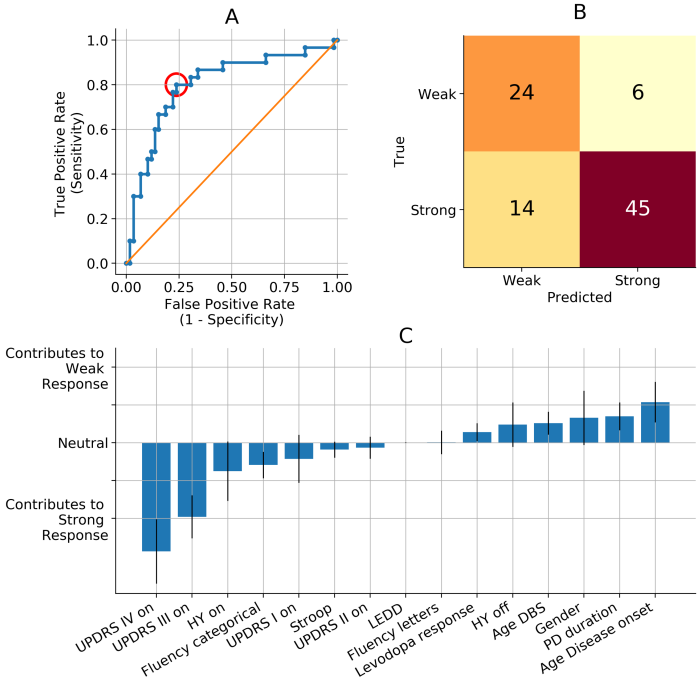


Figure 3: Prediction model performance and importance per predictive variable.
(A) Visualization of the performance of our prediction model. Our prediction model performs with an average area under the curve (AUC) of the receiver operating curve (ROC, blue line) of 0.78 (standard deviation: 0.08). All the dots on the ROC represent a threshold between 0 and 1 for accepting a probability to be a weak responder to be true. Every threshold leads to a different true positive rate and false positive rate. The red circle represents the threshold corresponding with B. The orange line represents chance level in which true positive rates equal true negative rates. (B) Confusion Matrix of the example when 0.29 is chosen as a threshold for accepting the probability to be a weak responder (red circle in A). The true positive rate of 0.80 results in 24 out of 30 true weak responders getting a true weak prediction. The false positive rate of 0.24 results in 14 out of 59 true strong responders getting a false weak prediction. The classification accuracy is 0.78 with 69 out of 89 correct predicted patients. (C) Relative influence of all preoperative predictive variables. The blue bars represent the normalized Odds Ratios. The heights represent the effect on prediction outcome

of a 1 unit increase in the specific variable, while all other variables stay equal. AUC, area under the curve; DBS, deep brain stimulation; H&Y, Hoehn & Yahr scale; LEDD, levodopa equivalent daily dosage; Levodopa response, difference between UPDRS III off-medication minus UPDRS III on-medication; off, off-medication; on, on-medication; ROC, receiver operate characteristic; TEED, total electrical energy delivered; UPDRS, Unified Parkinson Disease Rating Scale; PD, Parkinson's disease.

Performance of the prediction model

The prediction model has a good general performance with an average AUC of the ROC of 0.79 (standard deviation: 0.08) (figure 3A).

When 0.29 is chosen as a threshold for accepting probabilities to become a weak responder, this leads to a true positive rate of 0.80 and a false positive rate of 0.24 (fig. 3A-B). This corresponds to a positive predictive value of 0.63 and a negative predictive value of 0.88. Selecting 0.29 as the threshold for probability acceptance leads to a classification accuracy of 78%, since 69 out of 89 patients are predicted correctly.

The relative influence values represent the influence, or weight, of each preoperative variable in the prediction model (fig. 3C). Older age at PD onset has the strongest relative influence for becoming a weak responder. High preoperative UPDRS III and IV scores in the on-medication condition are the strongest predictors for becoming a strong responder (figure 3C). Additionally, a high preoperative UPDRS II score, high scores on the categorical fluency and Stroop interference test, and higher H&Y score in the on-condition were moderate predictors for becoming a strong responder.

Discussion

Proof of concept of machine learning prediction in preoperative DBS outcome counselling

The presented machine learning model differentiated between individual weak and strong motor responders one-year after STN DBS for PD with a good overall predictive performance, the AUC of the ROC was 0.78, and the classification accuracy was 0.78% (fig. 3A-B). These results contribute to a proof-of-concept of machine learning prediction of individual postoperative motor outcome, solely based on preoperative clinical variables. We want to stress that these results are the first step towards the clinical utilization of smart supportive computational models in the delicate, multifactorial decision-making process of DBS therapy counselling. To increase the likelihood of creating a beneficial clinical impact for the patient, a model should be interpretable for clinicians, generalizable over the aimed patient population, and the effect of a utilization on the quality of clinical care should be investigated.[28]

Interpretation of the predictive performance and the clinical utilization

A predictive machine learning model for clinical support generates individual outcome probabilities range from 0 to 1, rather than binary classes. The presented confusion matrix is an example of a clinical utilization where probabilities to become a weak responder higher than 0.29 were accepted (see fig. 3A-B). The selection of this threshold will eventually determine the model's clinical behaviour, usefulness, and its potential clinical impact on patient care. The value of this threshold leads to a different balance between false positive and false negative predictions (fig. 3B), and should be validated on an external cohort. The presented threshold is chosen to realize a good accuracy (78%) and to fit to the intended clinical utilization of this model. Since the

majority of STN DBS candidates will experience a strong response, it is important that the clinician can trust a strong response prediction (negative predictive rate (0.88). Also, the model should create awareness about the chance of becoming a weak responder in case of increased risk. This requires a true positive rate, here 0.80.
Further, the confusion matrix shows that most incorrect predictions are actual strong responders who get a weak responder prediction. The final decision will be accurately guided by the experience of the DBS team and will overrule the majority of these predictive inaccuracies. Therefore, the actual clinical usefulness and impact should be investigated in a prospective clinical study. Moreover, these numbers and considerations emphasize that a clinical decision support tool in a precarious setting as preoperative counselling for DBS therapy should have a warning role, instead of a directive role. The goal should be to support the clinician with validated numerical expectations, and ensure her or his awareness in case of a patient with a higher than average chance on suboptimal therapeutic effect.

The clinical value of predicting STN DBS motor response in the preoperative phase
Establishing an accurate prediction tool for motor outcome after STN DBS facilitates the clinician to improve patient counselling, expectation management, postoperative patient satisfaction, and potentially even patient selection.[29] Due to the complexity and heterogeneity of individual STN DBS candidates, outcome prediction needs to be accompanied by a clinical expert's appraisal. Moreover, the accuracy of a prediction model solely regarding clinical preoperative factors will always be limited due to the influence of surgical factors. Nevertheless, we intentionally chose to leave pre-, intra- and postoperative imaging and neurophysiology variables out of our model. This way, we ensure the model's accessibility and usability in clinical practice. We aim to provide the clinician during preoperative counselling with numerical support regarding the most probable motor outcome for an individual patient.
Further it is important to underline this model's target patient population and clinical utilization. The model is designed for, and tested on, PD patients who were included for STN DBS implementation. This means the model should be applied to patients which are highly likely to be included for STN DBS implementation in the current care practices. In this population, the model is aimed to inform the clinician, and indirectly the patient, about a potential increased risk on a suboptimal motor response. This means the model is not developed to identify optimal STN DBS candidates from a general PD population.

The additive value of machine learning methods for clinical decision support tools
The applied predictive multivariate logistic regression model was chosen to overcome limitations inherent to conventional (univariate) logistic regression models.[8,10,11] Traditional predictive or correlative analyses mainly result in a correlation between one preoperative variable and a postoperative outcome, while controlling for several confounding preoperative variables. The absence of confounders and predictive variable selection in machine learning models, makes them less limited by sample size than traditional correlative analyses.[25] The presented prediction model distinguishes itself by evaluating all available variables simultaneously. The applied cross-validation decreases the restriction due to sample size and leads to less a-priori selection-bias. Nevertheless, the advantages of machine learning predictive models come with specific analysing risks. For example, an external validation is required to evaluate under- or overfitting of the model, and validation of the threshold for accepting probabilities. Further, we stress the importance of using interpretable predictive machine learning models. In contrast to more

complicated models such as deep neural networks, interpretable machine learning models remain explainable. This is essential in evaluating clinical validity and creating clinical confidence in a supportive decision tool which are both important in realizing actual clinical impact.[26]

Interpretation of the preoperative predictive variables in this model

This overview of interpretable weight of each predictive variable is an advantage of the applied logistic regression in the prediction model (fig. 3C). This advantage enables clinicians to verify whether the ratio behind the predictions is clinically valid or whether predictions are based on unexpected variables.

The reported large influence of higher age at PD onset on becoming a weak responder contradicts a finding of a meta-analysis that report younger age to be a positive predictor for a favourable outcome. Contrarily, the same meta-analysis reports longer PD duration as a predictor for favourable outcome.[7]

Preoperative UPDRS III and IV scores in the on-medication condition have the largest relative influence values for becoming a strong responder in this model. High preoperative motor severity increasing the chance to become a strong responders is in line with the findings of a meta-analysis, although most included studies describe preoperative severity in the off-medication condition.[7] Evidence on the predictive value of symptom severity in on-medication condition is limited. The finding that H&Y scores do not majorly influence outcome probabilities is in line with previous literature. This literature describes that disease severity positively influences the chance on strong motor response, while axial and balance problems negatively influence this chance.[7] Since H&Y severity is based on both these factors, an inconclusive effect is expected.

Furthermore, there is literature on predictive or correlative variables and QoL outcome after STN DBS. We cannot compare these findings one-on-one with our findings. However, our holistic outcome classification aims to cover multiple aspects which influence QoL. High preoperative UPDRS III scores, and high UPDRS III levodopa response, are identified as important predictors of good QoL outcome, and motor outcome.[7,10,11] Conversely, recent studies have failed to replicate this positive predictive value of UPDRS III severity, or UPDRS III levodopa response on QoL outcome, or motor outcome.[8,9,12] Our findings are in line with some of these findings, since the absolute UPDRS III score showed a relevant influence, while the UPDRS III difference between on- vs. off-medication condition did not have a relevant influence. Regarding the reported influence of levodopa responsiveness, one should consider that LEDD is expressed in milligrams, which means that the relative influence of a unit increase (1 milligram) is not a clinically relevant increase.

A high score on categorical Fluency is a small contributor to becoming a strong responder (fig. 3C). A high categorical Fluency score corresponds to better neuropsychological functioning. The contribution of the Stroop interference score is very small. Thus, there is no large influence of neuropsychological tests in our prediction model.

Our lack of QoL scores prevented replication of previous findings which suggested that an impaired preoperative QoL-functionality predicts a large postoperative QoL improvement.[8,14] Likewise, the absence of a proper non-motor symptom scale hampered potential reproduction of the recently described importance of non-motor symptoms.[13]

The reported influences of the preoperative variables on the outcome probability are mainly consistent with the literature, and are partly contradicting literature. We stress that the reported

influences of this model cannot be seen outside the scope of this model. They are only reported to gain insight in the underlying weights which determine the probabilities.
They cannot be interpreted on their own within individual patients when other variables in the model are disregarded.

Limitations
Our study is limited by its retrospective character. Missing preoperative data points were overcome by imputations. Outcome values were not imputed to prevent training of the model based on imputed self-generated data. Even though the imputation method was sound, the imputed values will never reach true values and will influence outcomes. Second, the internal consensus on the applied categorization for motor outcome is based on scientific grounds, but can always be disputed. The holistic approach including UPDRS II, III and IV, aims to cover aspects of daily life activities, motor symptoms, and adverse effects of treatment. Future work should include QoL metrics and investigate the correlation between (QoL) and the presented classification. We argue our approach in the Supplementary Material. Lastly, the accuracy of a preoperative prediction model will always be limited and contain variance due to the lack of surgical variables.

Conclusion
The presented prediction model identified strong vs. weak responders one-year after STN DBS for PD with a good classification accuracy. The potential distribution of predictive inaccuracies was in line with the aimed clinical utilization. These findings contribute to the proof-of-concept of machine learning prediction of individual motor outcome after STN DBS based on preoperative clinical variables.
The reported preoperative variables cannot be interpreted separately outside the scope of this prediction model, but endorse the clinical reliability of the applied method.
These results and considerations support the potential and the timely relevance of predictive clinical support tools for DBS outcome, and advocate further reproduction and validation in a representative, multicenter cohort. The optimal clinical utilization should be refined and the clinical additional value and impact should be clarified before a predictive clinical support tool can be applied during individual preoperative DBS counseling.

Acknowledgements
We would like to thank Jackson Boonstra for proofreading the manuscript.

Supplementary material

Surgical procedure

PD patients were indicated for STN DBS based on severe motor symptoms despite optimal levodopa treatment, severe motor fluctuations, or dyskinesia, and often showed a good levodopa responsiveness.[30] Surgical electrode location was determined based on preoperative MRI trajectory planning, microelectrode recordings, and intra-operative testing. The first part of surgery was performed while the patient was cognizant, after which the lead and pacemaker implantation was completed under general anaesthesia. Postoperative CT examinations verified the electrode location. Postoperative stimulation parameters and dopaminergic drug therapy were set and managed by the neurologist in the outpatient clinic during regular follow-ups.

The applied categorization for postoperative motor response

Quantifying motor response after STN DBS is a challenging task at its own due to the variety in symptoms and the variety in decisive reasons for surgical inclusion. Symptom severity in on-medication state, time spend in off-time, frequency of on- vs. off-fluctuations, dopaminergic adverse effects all could be reasons to include a PD patient for STN DBS. Obviously, successful motor response after STN DBS is as diverse as these different reasons for surgery. Previous literature describes STN DBS outcome with a heterogenous variety of variables including UPDRS III severity in on- or off-conditions, time spent in on-condition, and QoL scales.[2,3,5,8]

To identify the suboptimal responding minority of STN DBS patients objectively, we pragmatically created a holistic categorization based on available variables representing ADL, motor symptoms, and adverse effects (paper fig. 2). The applied cut off values are based on existing literature and are argued in the next paragraph of these supplementary material. In the absence of a validated QoL scale, we consider our categorization as the best possible definition of general unsatisfactory outcome.

This categorization resulted in one-third weak responders, which is comparable to reported improvement ratios on quality of life after STN DBS.[5,10,13,14] Especially since we regarded postoperative differences in on-medication conditions instead of off-medication conditions; and postoperative surgical results are often comparable to the best state in on-medication condition.[7]

The strong and weak responders significantly differed on all UPDRS changes, except for UPDRS III scores in on-stimulation and off-medication condition compared to preoperative off-medication condition (table S1). A plausible explanation is that nearly all patients will benefit from stimulation compared to the untreated preoperative condition. We consider comparing on-medication states as more natural and realistic because therapy aims to keep patients in on-medication state most of the time.

Given the concordance regarding preoperative and postoperative UPDRS scores and LEDD amounts between our cohort and the literature, we consider our cohort to be representative for the general STN DBS population.[2,5] Our cohort showed an expected older age and more severe pre- and postoperative characteristics then a cohort included based on early motor complications.[4] The observed neuropsychological deteriorations after STN DBS are also in line with previous findings.[31]

Literature on clinically relevant UPDRS differences

We searched PubMed using search terms: (Parkinson*) AND (UPDRS) AND ((clinical* relevant) OR (clinical* significant)) AND (improve* OR change OR difference). We selected papers which aimed to define minimal clinically relevant UPDRS changes and included predictive papers which used absolute thresholds for UPDRS changes in defining outcome.

Table 1 shows the reported UPDRS differences suggested to be clinically relevant.

Reference	Compared (MDS-) UPDRS score	Sign difference
Kostoglu[32]	III Med-OFF vs. Med-OFF + STIM	38%
Horvath[22]	III	- 3.25 + 4.63
Shulman[33]	III Total	+/- 2.5 (minimal) +/- 5.2 (moderate) +/- 10.8 (large) +/- 4.3 (minimal) +/- 9.1 (moderate) +/- 17.1 (large)
Schrag[34]	III II (H&Y > 2)	+/- 5 3
Makkos[35]	II + III I + II + III Total	-4.9 + 4.2 -6.7 +5.2 -7.1 +6.3

Table S1: overview clinically significant changes in (MDS-)UPDRS scores reported in literature.

For the UPDRS III score, the relevant differences range between 2.5 and 5 for the smallest relevant change. In one predictive analysis, a cut off change of 38% was used for UPDRS III off-medication. This translates to an absolute change of >10 in many patients. We averaged these findings and set the cut off for clinically relevant improvement on UPDRS III at 5 points.
Since the UPDRS II and IV scores consist of fewer points, it is logical that the cut off values are lower. Based on the mentioned cut offs in the literature, we set the cut off for the UPDRS II and IV scores at 3 points. Every patient who improved more than 3 points on UPDRS II or IV was evaluated on UPDRS III change. When a patient deteriorated more than 7 points on UPDRS III, while improving on UPDRS II or IV, he or she was classified as a Weak responder. The 7 points cut off on UPDRS III represents the minimal clinical important difference of 5 points, and the natural disease progression of 2 points on UPDRS III scale per year[23].

Imputation methods
We imputed missing values in the presurgical parameters using Random Forrest Regressors or Classifiers for each individual parameter. The models were trained on all participants that had the data available while missing values were imputed based on the available presurgical parameters for the remaining participants. For continuous parameters, we used Random Forrest Regressors while categorical parameters were imputed using Random Forrest Classifiers. Only presurgical parameters were imputed to ensure that our results were not based on imputed values.

References

1 Limousin, P. et al. Effect of parkinsonian signs and symptoms of bilateral subthalamic nucleus stimulation. Lancet 345, 91-95 (1995).

2 Deuschl, G. et al. A randomized trial of deep-brain stimulation for Parkinson's disease. The New England journal of medicine 355, 896-908, doi:10.1056/NEJMoa060281 (2006).

3 Odekerken, V. J. et al. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. The Lancet. Neurology 12, 37-44, doi:10.1016/s1474-4422(12)70264-8 (2013).

4 Schuepbach, W. M. et al. Neurostimulation for Parkinson's disease with early motor complications. The New England journal of medicine 368, 610-622, doi:10.1056/NEJMoa1205158 (2013).

5 Williams, A. et al. Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. The Lancet. Neurology 9, 581-591, doi:10.1016/s1474-4422(10)70093-4 (2010).

6 Pinter, M. M., Alesch, F., Murg, M., Helscher, R. J. & Binder, H. Apomorphine test: a predictor for motor responsiveness to deep brain stimulation of the subthalamic nucleus. Journal of neurology 246, 907-913 (1999).

7 Kleiner-Fisman, G. et al. Subthalamic nucleus deep brain stimulation: summary and meta-analysis of outcomes. Movement disorders : official journal of the Movement Disorder Society 21 Suppl 14, S290-304, doi:10.1002/mds.20962 (2006).

8 Schuepbach, W. M. M. et al. Quality of life predicts outcome of deep brain stimulation in early Parkinson disease. Neurology, doi:10.1212/wnl.0000000000007037 (2019).

9 Zaidel, A., Bergman, H., Ritov, Y. & Israel, Z. Levodopa and subthalamic deep brain stimulation responses are not congruent. Movement disorders : official journal of the Movement Disorder Society 25, 2379-2386, doi:10.1002/mds.23294 (2010).

10 Daniels, C. et al. Is improvement in the quality of life after subthalamic nucleus stimulation in Parkinson's disease predictable? Movement disorders : official journal of the Movement Disorder Society 26, 2516-2521, doi:10.1002/mds.23907 (2011).

11 Frizon, L. A. et al. Quality of Life Improvement Following Deep Brain Parkinson's Disease: Development of a Prognostic Model. Neurosurgery, doi:10.1093/neuros/nyy287 (2018).

12 Abboud, H. et al. Predictors of Functional and Quality of Life Outcomes following Deep Brain Stimulation Surgery in Parkinson's Disease Patients: Disease, Patient, and Surgical Factors. Parkinsons Dis 2017, 5609163, doi:10.1155/2017/5609163 (2017).

13 Dafsari, H. S. et al. Short-term quality of life after subthalamic stimulation depends on non-motor symptoms in Parkinson's disease. Brain stimulation, doi:10.1016/j.brs.2018.02.015 (2018).

14 Liu, F. T. et al. Predictors to quality of life improvements after subthalamic stimulation in Parkinson's disease. Acta neurologica Scandinavica, doi:10.1111/ane.13056 (2018).

15 Goetz, C. G. et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement disorders : official journal of the Movement Disorder Society 23, 2129-2170, doi:10.1002/mds.22340 (2008).

16 Meyer, A. et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. The Lancet. Respiratory medicine 6, 905-914, doi:10.1016/s2213-2600(18)30300-x (2018).

17 Pieter Kubben, M. D., Andre Dekker. Fundamentals of Clinical Data Science. 1 edn, v-vi (Springer International Publishing, 2019).

18 Cerasa, A. Machine learning on Parkinson's disease? Let's translate into clinical practice. J Neurosci Methods 266, 161-162, doi:10.1016/j.jneumeth.2015.12.005 (2016).

19 Ballarini, T. et al. Regional gray matter changes and age predict individual treatment response in Parkinson's disease. NeuroImage. Clinical 21, 101636, doi:10.1016/j.nicl.2018.101636 (2019).

20 Esselink, R. A. et al. Unilateral pallidotomy versus bilateral subthalamic nucleus stimulation in PD: a randomized trial. Neurology 62, 201-207 (2004).

21 Koss, A. M., Alterman, R. L., Tagliati, M. & Shils, J. L. Calculating total electrical energy delivered by deep brain stimulation systems. Annals of neurology 58, 168; author reply 168-169, doi:10.1002/ana.20525 (2005).

22 Horvath, K. et al. Minimal clinically important difference on the Motor Examination part of MDS-UPDRS. Parkinsonism & related disorders 21, 1421-1426, doi:10.1016/j.parkreldis.2015.10.006 (2015).

23 Holden, S. K., Finseth, T., Sillau, S. H. & Berman, B. D. 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 5, 47-53, doi:10.1002/mdc3.12553 (2018).

24 Pedregosa, F. et al. Scikit-learn: Machine learning in Python. Journal of machine learning research 12, 2825-2830 (2011).
25 Wynants, L. et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. Journal of clinical epidemiology 68, 1406-1414, doi:10.1016/j.jclinepi.2015.02.002 (2015).
26 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206 (2019).
27 Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC medicine 13, 1 (2015).
28 Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17, 195, doi:10.1186/s12916-019-1426-2 (2019).
29 Lin, H. Y., Hasegawa, H., Mundil, N., Samuel, M. & Ashkan, K. Patients' Expectations and Satisfaction in Subthalamic Nucleus Deep Brain Stimulation for Parkinson Disease: 6-Year Follow-up. World Neurosurg 121, e654-e660, doi:10.1016/j.wneu.2018.09.181 (2019).
30 Temel, Y. et al. Single electrode and multiple electrode guided electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease. Neurosurgery 61, 346-355; discussion 355-347, doi:10.1227/01.neu.0000303993.82149.98 (2007).
31 Okun, M. S. Deep-brain stimulation for Parkinson's disease. The New England journal of medicine 367, 1529-1538, doi:10.1056/NEJMct1208070 (2012).
32 Kostoglou, K. et al. Classification and Prediction of Clinical Improvement in Deep Brain Stimulation From Intraoperative Microelectrode Recordings. IEEE transactions on bio-medical engineering 64, 1123-1130, doi:10.1109/tbme.2016.2591827 (2017).
33 Shulman, L. M. et al. The clinically important difference on the unified Parkinson's disease rating scale. Archives of neurology 67, 64-70, doi:10.1001/archneurol.2009.295 (2010).
34 Schrag, A., Sampaio, C., Counsell, N. & Poewe, W. Minimal clinically important change on the unified Parkinson's disease rating scale. Movement disorders : official journal of the Movement Disorder Society 21, 1200-1207, doi:10.1002/mds.20914 (2006).
35 Makkos, A. et al. Are the MDS-UPDRS-Based Composite Scores Clinically Applicable? Movement disorders : official journal of the Movement Disorder Society 33, 835-839, doi:10.1002/mds.27303 (2018).

Chapter 3

Multi-center validation of DBS-PREDICT – Preoperative prediction of motor outcome after subthalamic deep brain stimulation in Parkinson's disease

Habets JGV, Herff C, Fasano AA, Beudel M, Kocabicak E, Schnitzler A, Abu Snineh M, Kalia SK, Ramirez C, Hodaie M, Munhoz RP, Rouleau E, Yildiz O, Linetsky E, Schuurman R, Hartmann CJ, Lozano AM, De Bie RMA, Temel Y, Janssen MLF

Under review, May 2021

Abstract

Background: Subthalamic deep brain stimulation (STN DBS) is an established therapy for Parkinson's disease (PD) patients suffering from motor response fluctuations despite optimal medical treatment, or severe dopaminergic side effects. Despite careful clinical selection and surgical procedures, some patients do not benefit from STN DBS. Preoperative prediction models are suggested to better predict individual motor response after STN DBS. We validate a preregistered model, DBS-PREDICT, in an external multicenter validation cohort.

Methods: DBS-PREDICT considered eleven, solely pre-operative, clinical characteristics, and applied a logistic regression to differentiate between weak and strong motor responders. Weak motor response was defined as no clinically relevant improvement on UPDRS II, III or IV scales, one-year after surgery, defined as respectively 3, 5, and 3 points or more. Lower UPDRS III and IV scores, and higher age at disease onset contributed most to weak response predictions. Individual predictions were compared with actual clinical outcomes.

Results: 322 PD patients treated with STN DBS, from six different centers were included. DBS-PREDICT differentiated between weak and strong motor responders with an area under the receiver operator curve of 0.76, and an accuracy up to 77%.

Conclusion: Generalizability and feasibility of individual STN DBS outcome prediction based on preoperative variables are demonstrated in an external, multicenter cohort. For further development of clinically impactful prediction tools for individuals in DBS care, prospective studies are required to overcome several inherent practical and statistical limitations.

Introduction

Parkinson's disease (PD) is a neurodegenerative disease characterized by motor and non-motor symptoms and will affect up to 12 million patients worldwide by 2040.[1, 2] Oral dopamine replacement therapies treat early disease symptoms well, disease progression or wearing-off effects may cause motor and non-motor fluctuations.[3] Subthalamic nucleus (STN) deep brain stimulation (DBS) is an accepted and proven symptomatic treatment of early and advanced motor response fluctuations with satisfactory overall improvement in the majority of patients.[4-6] However, up to one-third of patients experiences suboptimal motor improvement postoperatively,[7, 8] and decreasing this amount potentially would have to lead to clinical and socioeconomic impact. To realize this, several strategies are suggested such as improving preoperative selection,[9, 10] optimizing surgical lead placement,[11] and optimizing postoperative DBS programming or stimulation paradigms.[12-14]

To improve preoperative selection, the relation between preoperative variables and postoperative outcome has to be understood better. Preoperative significant quality of life (QoL) impairment, severe motor symptoms, greater levodopa response, better cognitive performance, tremor-dominancy, and younger age, have been correlated with better postoperative motor function improvement and QoL.[15-19] It is important to note that there is no consensus about all relevant preoperative predictors, and some of them specifically are a matter of debate.[20] Unlike the studies reporting these correlations,[15-18] two studies developed a machine learning model to decrease the number of suboptimal responders with individual motor response predictions (figure 1).[21, 22] Both proof-of-concept studies described a good prediction of a binary outcome in a retrospective single-center STN DBS cohort. They both were designed for preoperative application and did not consider intra- or post-operative variables.

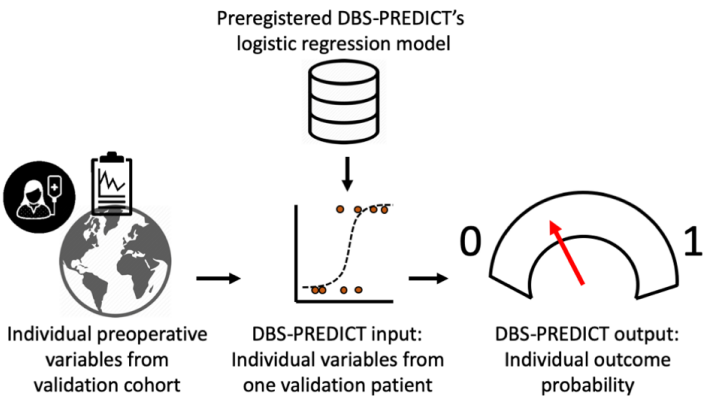


Figure 1: Schematic representation of methodology. Individual preoperative variables from six international centers are used as validation cohort (left panel). DBS-PREDICT's preregistered logistic regression model is trained on a separate development cohort which is described earlier. It takes preoperative variables from one individual patient as input (middle panel). DBS-PREDICT returns an individual probability for showing a weak motor response one-year after STN DBS (right panel).

Frizon et al defined favorable outcome as clinically relevant QoL improvement, and their model considered three preoperative variables which were most associated with favorable outcome. A logistic regression model was applied and validated using bootstrap sampling.[21] The model applied in this study, DBS-PREDICT, defined unfavorable outcome as no clinically relevant improvement on the Unified Parkinson's Disease Rating Scale (UPDRS) II, III and IV one-year after DBS implantation (figure 2). This outcome classification is strict, as it only requires improvement

on one of the three variables to claim favorable outcome. DBS-PREDICT also utilized a logistic regression model which considered all 15 available preoperative variables and was validated via a ten-fold cross-validation.[22] To enable model generalizability and statistical transparency, DBS-PREDICT was retrained without preoperative neuropsychological and UPDRS I scores, and preregistered before data collection completion (ClinicalTrials.gov, NCT04093908).[23] Performance and predicting variable importance did not change relevantly after retraining (figure 3 and supplementary material).

To develop valid and generalizable clinical prediction models with an impactful clinical utilization, external validation is needed after model development and proof of concept.[24, 25] Here, we present an external multicenter validation of the preregistered prediction model DBS-PREDICT. Additionally, we discuss next steps towards clinical impactful tools for individual DBS outcome prediction.

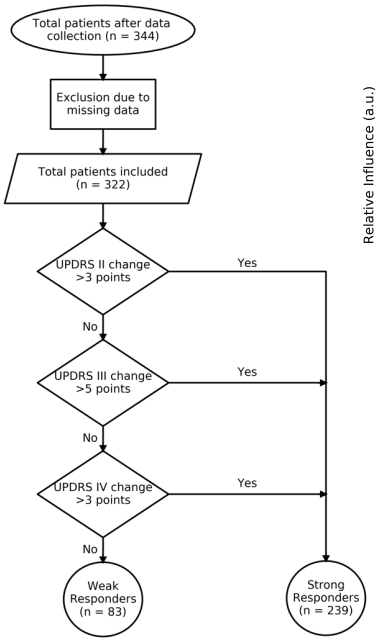


Figure 2 (left): Flowchart demonstrating the definition of strong and weak motor response. Only if no clinically relevant improvement in UPDRS II, III, or IV is found, a patient is considered to be a weak responder. UPDRS: Unified Parkinson Disease Rating Scale.

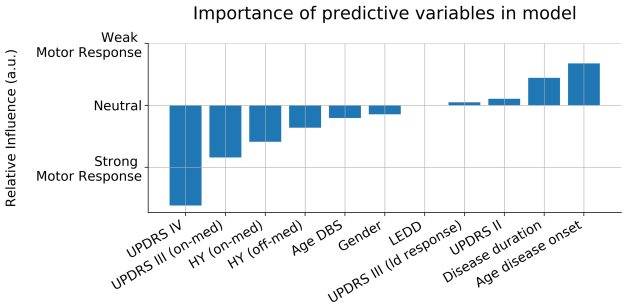


Figure 3 (right): Importance of all available preoperative predictive variables in DBS-PREDICT's logistic regression analysis. Importances are expressed in normalized relative weights. Relative weights are retrieved from DBS-PREDICT's logistic regression function after training the model on the development cohort.

Methods

DBS-PREDICT development

Development details of the model are described previously,[22] and will be briefly summarized here. In the development study, all available preoperative predictors were included, without preselecting predictors based on a sensitivity analysis. The preoperative variables included are gender, age at PD onset, PD disease duration, age at DBS, levodopa equivalent daily dosage (LEDD), UPDRS I, II, III, and IV, Hoehn & Yahr (H&Y) scales, and relevant neuropsychological assessments (Stroop test interference and Verbal Fluency categorical and letter tests). All

preoperative tests were reported in on-medication condition, only UPDRS III and H&Y scale were reported both in on- and off-medication condition.
Strong motor response was defined as a minimal clinical important difference (MCID) on UPDRS II, or III, or IV one year after STN DBS implementation (figure 2). Based on literature, MCID's were set as respectively 3, 5, and 3 points.[34-38] We chose to consider only UPDRS III change in on-stimulation and on-medication condition, compared to preoperative on-medication because the daily life reflection of the fully dopamine-deprived off-medication condition can be disputed. The daily life influence of off-medication moments, we aim to capture via the narrative items regarding motor aspects of experiences of daily living in UPDRS II, and regarding off-fluctuations in UPDRS IV. This definition results in weak responders who do not show a MCID in activities of daily life (UPDRS II), nor in motor symptoms in optimal therapy condition (UPDRS III), nor in the burden or duration of adverse effects and off-periods (UPDRS IV).
The proof of concept analysis demonstrated an area under the curve (AUC) of the receiver operator curve (ROC) of 0.88 (+/- 0.14), and a classification accuracy of 78%.[22] In this analysis, optimal predictive performance was achieved when generated probabilities to become a weak responder above 0.24 were accepted as indicative. The importance of preoperative variables driving the prediction, and the classification accuracy did neither show significant, nor relevant change after the exclusion of these variables (figure 3 and supplementary material). The model is made available on GitHub (https://github.com/jgvhabets/DBSPREDICT).

Patient population
A retrospective multi-center cohort was collected as validation dataset. DBS centers were invited to provide data from PD patients who were treated with STN DBS. The following preoperative variables were collected: age at PD onset, age at DBS, disease duration, gender, LEDD, preoperative UPDRS II, III in on-medication and off-medication conditions, and IV scores, and HY scales in on- and off-medication conditions. The following postoperative variables were collected from one-year follow-up moments: UPDRS II, III in on-medication and on-stimulation conditions, and IV scores.
Approval of a local ethical committee was received that this study agreed with the Good Clinical Practice norms and the Declaration of Helsinki (METC azM/UM: 2018-0739-A-9).

Descriptive statistical methods
Pre- and postoperative variables from the current multi-center population and the initial single-center training population were compared. Continuous variables were evaluated for significant differences with Mann-Whitney-U analyses since sample sizes differed and not all variables were normally distributed. The distributions of weak responders in the multi-center validation cohort and the development cohort were compared using a Chi-squared test.

Predictive statistical methods
The model generated a probability to become a weak responder based on the preoperative variables for every individual patient. If the generated probability is higher than a preset threshold, the model predicts a weak motor response. The predictive performance was analyzed with both the default threshold of 0.5,[26] as well as the optimal threshold from the development study, 0.24, since the development cohort was as unbalanced as the validation cohort.[22] To provide a valid and understandable evaluation of the performance of the model, we report the

AUC of the ROC, positive and negative predictive values (respectively PPV and NPV), the sensitivity and the accuracy.
Findings are presented in line with TRIPOD guidelines for artificial intelligence predictive modeling.[39, 40]

Results

Patient population

Six different DBS centers in Europe, Asia and North America participated in this validation study. In total, 344 PD patients were included who were treated with STN DBS, completed a one-year postoperative follow up, and whose records contained all required variables. After removing missing data, 322 patients were included in the validation cohort. The descriptive statistics and distributions of the pre- and postoperative variables of the validation cohort are shown and compared with the development cohort in supplementary table 1**.**

The presented validation cohort had 26% weak responders (83 out of 322), compared to 30% in the development cohort (27 out of 90). This indicated that the populations do not have statistically significant differences in distribution of strong and weak responders (Chi-square statistic of 2.892, p = 0.089). To test statistical significance of individual pre- and postoperative variables, a Bonferroni corrected p-value of 0.0038 is used. The validation cohort had a lower rate of female patients. The preoperative UPDRS III scores were significantly higher in the development cohort, together with a significantly lower LEDD preoperatively. Postoperatively, both UPDRS III and LEDD did not differ significantly between the two cohorts.

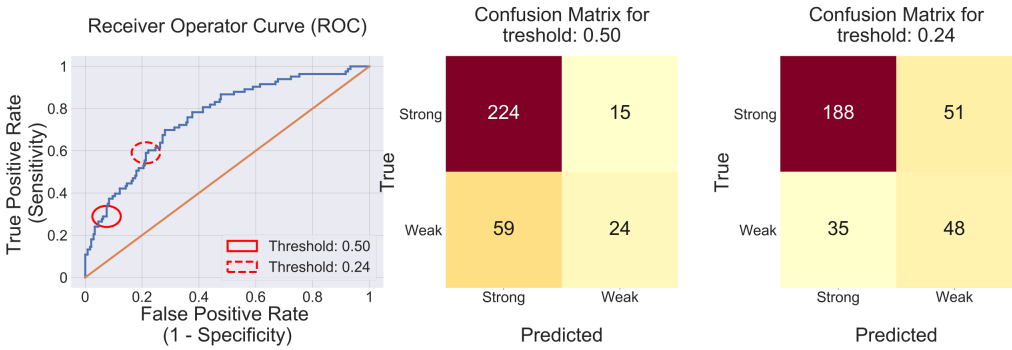


Figure 4: Performance of prediction model in validation cohort. Left: receiver operator curve showing true positive rates and false positive rates for different thresholds, both analyzed thresholds are showed with red circles. Middle: confusion matrix corresponding to 0.50 threshold. Right: confusion matrix corresponding to 0.24 threshold.

Predictive performance

The individual preoperative variables of all 322 were inserted in DBS-PREDICT, which led to 322 binary outcome predictions. This was repeated with two different scenarios. In the first scenario, a default threshold for probability acceptance of 0.50 was chosen.[26] In the second scenario, the optimal threshold from the development study, 0.24, was applied.[22] The generated binary outcome predictions were compared with the binary outcome classes based on the actual postoperative variables.

DBS-PREDICT differentiated between strong and weak motor responders with an AUC of 0.76 (figure 4, left panel). The default threshold classified all probabilities above 0.5 as weak predictions (figure 4 middle panel). This led to an accuracy of 0.77 (248 correct predictions, out of 322). The numbers in the confusion matrix led to a sensitivity of 0.29 (predicted 24 out of 83 true weak responders correctly), a positive predictive value of 0.62 (out of 39 weak predictions, 24 were correct), and a negative predictive value of 0.79 (out of 283 strong predictions, 224 were correct) (table 1). The threshold chosen after the results of the development cohort (0.24) (figure 4, right panel), led to an accuracy of 0.73 (236 correct predictions, out of 322). This threshold led to a sensitivity of 0.58 (predicted 48 out of 83 true weak responders correctly), a positive predictive value of 0.49 (out of 99 weak predictions, were correct 48), and a negative predictive value of 0.84 (out of 223 strong predictions, 188 were correct) (table 1).

	Threshold 0.5	**Threshold 0.24 ***
Accuracy	0.77 (248 / 322)	0.73 (236 / 322)
Sensitivity	0.29 (24 / 83)	0.58 (48 / 83)
Positive Predictive Value	0.62 (24 / 39)	0.49 (48 / 99)
Negative Predictive Value	0.79 (224 / 283)	0.84 (188 / 223)

Table 1: Predictive performance of DBS-PREDICT for two analyzed thresholds for accepting probabilities to become a weak responder. *: threshold from development results.

Discussion

We demonstrated the feasibility and generalizability of the preregistered DBS-PREDICT model in a retrospective multicenter validation cohort which is representative for the global STN DBS population. Our results are an important next step towards impactful clinical decision support systems (CDSS) for DBS care. They underline the potential of preoperative individual motor response prediction as a method to improve STN DBS motor outcome.[21, 22] In general, the translation from a potent machine learning prediction model into a CDSS with clinical impact is notoriously complex and faces several inherent challenges.[27, 28] Therefore, our findings and considerations are valuable for DBS CDSS development regardless of input variables or DBS indication.

Clinical impact of a CDSS is defined as the change in patient care resulting from its implementation,[24] and depends on the following three factors: 1) the generalizability of the model to external patient populations and centers, [24, 25] 2) the quality and understandability of numeric predictive results of a model,[24] and 3) the presence of a well-considered, clinically realistic utilization, including the socioeconomic impact of the utilization.[29] We will discuss how our results contribute to points 1 and 2, which limitations should be considered, and how this work paves the way for investigating point 3.[25]

Model generalizability and usability

Generalizability of DBS-PREDICT was demonstrated by the high accuracy and AUC score in the external validation cohort. This cohort was representative for the global STN DBS population.[4-6, 8] The small differences in preoperative characteristics between the development and validation cohort were not troublesome (supplementary table 1). The most relevant difference between the two cohorts were the higher preoperative UPDRS III scores in the development cohort together with the lower amount of LEDD (supplementary table 1). This could be explained by a more conservative pharmacological strategy in the development cohort compared to the validation

cohort. Since the postoperative separate UPDRS scores, the weak responder ratio, and the LEDD all not differed postoperatively, we assume both cohorts were comparable and representative STN DBS cohorts.
To ensure generalizability and usability among different DBS centers, DBS-PREDICT is deliberately founded on preoperative, clinical variables which do not require specific radiologic, or neurophysiological features. It is beyond doubt that perioperative and postoperative, surgical, radiological, and neurophysiological variables influence a patient's motor response. DBS-PREDICT aims nonetheless to provide a numeric support about expected motor response during preoperative counseling, within the possibilities of this conceptual limitation. Also, more detailed clinical variables such as UPDRS sub scores instead of sum scores have the theoretical potential to improve individualization of the prediction model. In a retrospective setting it is difficult to include large enough populations including similar detailed scores. Therefore, we suggest preregistered, prospective data collection to investigate the additional value of detailed clinical variables.

Quality, understandability and interpretation of predictive performance
DBS-PREDICT showed a good classification accuracy in both analyzed utilizations of 0.77 and 0.73, and an AUC of 0.76 (figure 4, table 1). The absence of a comparable validation study of a DBS outcome CDSS in the literature does not allow comparison of these results. However, results are in line with the two available model development studies.[21, 22]. This supports the feasibility of preoperative identification of weak motor responders in STN DBS care. Nevertheless, the clinical utilization of DBS-PREDICT still has to be discussed thoroughly. The current positive predictive values and sensitivities are not sufficient to withdraw predicted weak responders from DBS implantation (table 1). This emphasizes that a CDSS for a clinically highly complex and multi-factorial process such as DBS candidate selection should be approached with caution.[30] Clinicians need to understand the predictive performance of a CDSS and how it can support the DBS team, rather than replacing the team of clinicians in decision making.[24] The applied logistic regression, intentionally chosen over more complex machine learning models, contributes to this interpretability among clinicians. In general, logistic regression models are often non-inferior compared to more complex models for clinical applications.[31]
To improve both the predictive performance and the clinical understanding among clinicians, future CDSS development could combine clinical judgement and numerical support by a CDSS. For example, a CDSS could be designed and validated in a preoperative selected population by the clinician. This would also allow to explore the interaction between clinician and CDSS.

Tradeoffs and considerations of an impactful clinical utilization
As touched upon in the latter paragraph, a clinical utilization should be realistic to implement and be clinically useful and impactful. During the development and validation of DBS-PREDICT, all included patients underwent STN DBS based on the current clinical selection. To develop a CDSS to improve STN DBS inclusion, referred surgical candidates who are rejected in current practice should be included in future CDSS development and validation.
The exact moment of use in clinical DBS practice will also influence the interpretation of a CDSS' balance between false positive and false negative prediction. The consequences of a weak responder falsely predicted to respond strong, and vice versa, are depending on the CDSS' role in clinical decision-making. A broad understanding of these challenges among clinicians and

researchers is of major importance for the increasing attention CDSSs will get the upcoming decade.[24, 28, 30]

The potential socioeconomic impact

DBS care in PD is subject to large variations in disease symptomatology and progression, therapeutic strategies, and international differences in insurance and health care systems. These factors complicate a valid indication or measurement of the potential socio-economic impact of a DBS outcome CDSS. Current DBS therapy in PD is cost-effective compared to best medical treatment. This means the increased QoL is relatively larger than the increased health care costs.[32] Improved preoperative counseling and selection with CDSS's hold potential to lower absolute perioperative costs by preventing unsuccessful cases. Moreover, patients experiencing a less favorable outcome might be better prepared for this due to the improved preoperative counseling. This could lead to increased postoperative satisfaction, resulting in a better quality of life.[33] Thorough socioeconomic analyses should be included in future prospective CDSS development and data collection.[25] When a CDSS is not proven yet, data collection without interfering in clinical care should be considered.[25]

Alternative approaches for preoperative STN DBS outcome improvement

We want to stress that preoperative prediction based on clinical variables is one of several approaches to improve patient outcome. Advances in structural and functional imaging of the whole brain and its connectivity, hold potential to improve the preoperative target selection and surgical planning.[9] Better understanding of the STN itself can lead to more stable motor improvements as well.[11] Moreover, increasing insights in the clinical non-motor characteristics, genetic profiles, and neurophysiological characteristics and their relation with DBS outcome can improve patient selection, counseling and outcome.[18] All these approaches do not exclude each other and can improve DBS outcome in PD in a complementary way in the future. The discussed considerations regarding clinical impact realization, outcome definition and future study design are relevant for all different DBS CDSS's.

Limitations

This study is subject to several limitations, of which some are inherent to its retrospective nature. Limitations are minimized by preregistering the model, which mimicked a prospective application of the model. While the preregistration strengthened the statistical value, it did not allow to include new variables. We attempted to overcome the lack of a QoL-instrument as an outcome variable, by using a holistic outcome classification. The outcome classification covers minimal clinical important differences in a wide spectrum of factors influencing QoL: daily life performance, motor symptom severity, and adverse effects suffering (respectively UPDRS II, III, IV). To ensure generalizability, only sum scores were included instead of detailed clinical variables. Variance in clinical and administrative processes only allows a complete and detailed inclusion of pre- and postoperative variables via prospective data collection. A more complete and detailed data collection would require a development study first, and an external validation study afterwards.

The binary outcome classification can be disputed. This approach, however, aligns with the clinical decision in the individual patient to offer DBS or not. We emphasize that we followed large DBS trials in the pragmatic choice of selecting UPDRS sum scores as outcome variables.[7, 8] For the

evidence and considerations we provide here, consensus about the outcome definition does not have the highest priority. For future prospective, preregistered, DBS CDSS data collection, this has one of the highest priorities.
Local variation in pharmacological and DBS-programming strategies are neglected by the model. Due to the amount of data required by data-driven models, this currently cannot be overcome, and a wide generalizability is essential. Further, preoperative prediction of DBS outcome will always be limited by the absence of peri- and postoperative variables describing surgical electrode placement and surgical complications. This limitation is inherent to clinical role of a preoperative CDSS. Postoperative approaches to improve DBS outcome like directional steered stimulation or closed-loop DBS should always be considered for a sub optimal responding patient group.

Conclusion

Preoperative individual STN DBS outcome prediction is feasible in an external multicenter cohort using a preregistered model. The results and perspectives endorse next steps in the development of impactful CDSSs for individualized DBS care.
We emphasize the need of prospective and preregistered multicenter studies to overcome clinical, practical, and statistical challenges in the development of CDSS models for individual DBS decision making in PD care. These studies can be considered an obligatory next step before clinical implementation.

Supplementary Material

Literature for deciding minimal clinical important differences of outcome variables

A PubMed search was done using search terms: (Parkinson*) AND (UPDRS) AND ((clinical* relevant) OR (clinical* significant)) AND (improve* OR change OR difference). Papers were selected which aim to define minimal clinical important differences of UPDRS changes. The majority of evidence found describes UPDRS III differences. Table S1 summarizes results of this search.

	UPDRS scores subject of research	**Minimal clinical important difference**
Kostoglu[34]	III Med-OFF vs. Med-OFF + STIM	38%
Horvath[35]	III	- 3.25 + 4.63
Shulman[38]	III Total	+/- 2.5 (minimal) +/- 5.2 (moderate) +/- 10.8 (large) +/- 4.3 (minimal) +/- 9.1 (moderate) +/- 17.1 (large)
Schrag[36]	III II (H&Y > 2)	+/- 5 3
Makkos[37]	II + III I + II + III Total	-4.9 + 4.2 -6.7 +5.2 -7.1 +6.3

Table S1: overview clinically significant changes in UPDRS scores reported in literature.

For the UPDRS III score, the relevant differences range between 2.5 and 5 for the smallest relevant change. In one predictive analysis, a cut off change of 38% was used for UPDRS III off-medication. This translates to an absolute change of >10 in many patients. We averaged these findings and set the cut off for clinically relevant improvement after STN DBS at 5 points. Since the UPDRS II and IV scores consist of fewer points, it is logical that the cut off values are lower. Based on the mentioned cut offs in the literature, we set the cut off for the UPDRS II and IV scores at 3 points.

Adjustment prediction model

For feasibility reasons of the data collection, we trained our prediction model after excluding neuropsychological variables and without preoperative UPDRS I as preoperative predictors. The performance of the model did not change significantly. The area under the receiver operator curve was 0.85 (standard deviation 0.06), and the optimal classification accuracy was 0.83. The weights of the individual predicting variables were also comparable to the weights of the earlier published model.[22] The adjusted model was preregistered and not optimized anymore based on the newly collected data for the validation cohort.[23]

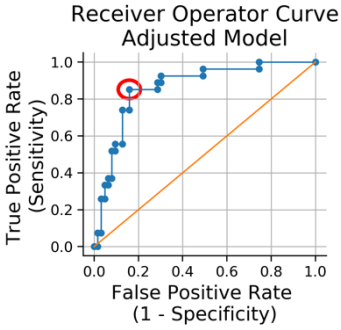


Figure S1: Receiver operator curve of adjusted model without neuropsychological variables and UPDRS I.

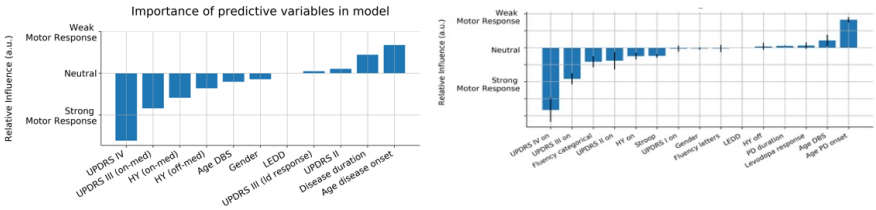


Figure S2 (left): Variable weights in adjusted model without neuropsychological variables and UPDRS I. Figure S3 (right): Variable weights in original model published in pre-print (adapted from Habets et al, medrXiv 2019)

	Current cohort (mean (standard deviation), unless †)	Development cohort (mean (standard deviation), unless †)
Total number of patients collected	334†	127 †
Included patients after exclusion due to missing data	322†	90 †
Gender (female/ male)	95/227 †	37/53 †,††
Age at DBS surgery (years)	0058 (8)	0061 (8)
Disease duration (years) at DBS surgery	0011.5 (4.7)	0010.6 (5.1)
LEDD preoperative (mg)	1525 (648)	1197 (622) ††
LEDD one year postoperative (mg)	0731 (444)	0665 (513)
UPDRS II preoperative*	0010.7 (6.9)	0009.8 (6.5)
UPDRS II one year postoperative*	0008.8 (5.4)	0009.7 (5.5)
UPDRS III preoperative*	0017.3 (8.7)	0021.8 (12.5) ††
UPDRS III preoperative, off-medication	0044.4 (14.3)	0039.4 (13.2) ††
UPDRS III preoperative change**	0 -27.1 (12.2)	0 -18.9 (13.4) ††
UPDRS III one year postoperative*	0015.1 (8.5)	0016.6 (10.0)
UPDRS IV preoperative*	0007.8 (3.3)	0005.5 (4.0)
UPDRS IV one year postoperative*	0003.5 (2.8)	0002.7 (2.4)
Preoperative HY scales* (in number of patients)	1: 01.9 %† 1.5: 02.2 % 2: 69.9 % 2.5: 17.1 % 3: 07.5 % 4: 01.2 % 5: 00.3 %	1: 02.4 % † 1.5: 02.4% 2: 16.5 % 2.5: 40.0 % 3: 28.2 % 4: 10.6 % 5: 000 %
Preoperative HY scales in off-medication (in number of patients)	1: 00.0 % † 1.5: 00.3 % 2: 37.0 % 2.5: 16.5 % 3: 33.9 % 4: 09.6 % 5: 02.8 %	NA

Table S1: Descriptive pre- and postoperative variables.
*: best medical condition, on-medication for preoperative variables, and on-medication and on-stimulation for postoperative variables; **: preoperative difference between off-medication vs. on-medication; †: numerical description, mean not applicable; ††: statistically significant difference between cohorts, p-value < 0.0038 (Bonferroni corrected for 13 variables). Gender compared with chi-squared test, other variables compared with Mann-Whitney-U test. DBS: deep brain stimulation, HY: Hoehn and Yahr, LEDD: levodopa equivalent daily dosage, mg: milligram, UPDRS: Unified Parkinson Disease Rating Scale.

References

1. Dorsey ER, Sherer T, Okun MS, Bloem BR. The Emerging Evidence of the Parkinson Pandemic. J Parkinsons Dis 2018;8(s1):S3-S8.
2. Titova N, Padmakumar C, Lewis SJG, Chaudhuri KR. Parkinson's: a syndrome rather than a disease? J Neural Transm (Vienna) 2017;124(8):907-914.
3. Kalia LV, Lang AE. Parkinson's disease. Lancet 2015;386(9996):896-912.
4. Schuepbach WM, Rau J, Knudsen K, et al. Neurostimulation for Parkinson's disease with early motor complications. The New England journal of medicine 2013;368(7):610-622.
5. Odekerken VJ, van Laar T, Staal MJ, et al. Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. Lancet Neurol 2013;12(1):37-44.
6. Janssen ML, Duits AA, Turaihi AH, et al. Subthalamic nucleus high-frequency stimulation for advanced Parkinson's disease: motor and neuropsychological outcome after 10 years. Stereotactic and functional neurosurgery 2014;92(6):381-387.
7. Williams A, Gill S, Varma T, et al. Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. The Lancet Neurology 2010;9(6):581-591.
8. Deuschl G, Schade-Brittinger C, Krack P, et al. A randomized trial of deep-brain stimulation for Parkinson's disease. The New England journal of medicine 2006;355(9):896-908.
9. Bari AA, Fasano A, Munhoz RP, Lozano AM. Improving outcomes of subthalamic nucleus deep brain stimulation in Parkinson's disease. Expert Review of Neurotherapeutics 2015;15(10):1151-1160.
10. Hartmann CJ, Fliegen S, Groiss SJ, Wojtecki L, Schnitzler A. An update on best practice of deep brain stimulation in Parkinson's disease. Therapeutic Advances in Neurological Disorders 2019;12:1756286419838096.
11. Bot M, Schuurman PR, Odekerken VJJ, et al. Deep brain stimulation for Parkinson's disease: defining the optimal location within the subthalamic nucleus. J Neurol Neurosurg Psychiatry 2018;89(5):493-498.
12. Picillo M, Lozano AM, Kou N, Puppi Munhoz R, Fasano A. Programming Deep Brain Stimulation for Parkinson's Disease: The Toronto Western Hospital Algorithms. Brain Stimul 2016;9(3):425-437.
13. Pina-Fuentes D, Little S, Oterdoom M, et al. Adaptive DBS in a Parkinson's patient with chronically implanted DBS: A proof of principle. Mov Disord 2017;32(8):1253-1254.
14. Schnitzler AB, M.; Verhagen, L.; Groppa, S.; Cheeran, B.; Karst, E.; Defresne, F.; Vesper, J. Directional versus omnidirectional Deep Brain Stimulation for Parkinson's disease: Results of a prospective, blinded, multi-center, single-arm crossover study [abstract]. Movement Disorders Society Conference. Nice, France2019.
15. Schuepbach WMM, Tonder L, Schnitzler A, et al. Quality of life predicts outcome of deep brain stimulation in early Parkinson disease. Neurology 2019.
16. Daniels C, Krack P, Volkmann J, et al. Is improvement in the quality of life after subthalamic nucleus stimulation in Parkinson's disease predictable? Mov Disord 2011;26(14):2516-2521.
17. Abboud H, Genc G, Thompson NR, et al. Predictors of Functional and Quality of Life Outcomes following Deep Brain Stimulation Surgery in Parkinson's Disease Patients: Disease, Patient, and Surgical Factors. Parkinson's disease 2017;2017:5609163.
18. Dafsari HS, Weiss L, Silverdale M, et al. Short-term quality of life after subthalamic stimulation depends on non-motor symptoms in Parkinson's disease. Brain Stimul 2018.
19. Cavallieri F, Fraix V, Bove F, et al. Predictors of Long-Term Outcome of Subthalamic Stimulation in Parkinson Disease. Annals of neurology.
20. Geraedts VJ, Feleus S, Marinus J, van Hilten JJ, Contarino MF. What predicts quality of life after subthalamic deep brain stimulation in Parkinson's disease? A systematic review. Eur J Neurol 2019.
21. Frizon LA, Hogue O, Achey R, et al. Quality of Life Improvement Following Deep Brain Parkinson's Disease: Development of a Prognostic Model. Neurosurgery 2018.
22. Habets JGV, Janssen MLF, Duits AA, et al. Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease—proof of principle in a retrospective cohort. PeerJ 2020;8:e10317.

23. ClinicalTrials.gov. NCT04093908: Prediction of STN DBS Motor Response in PD (DBS-PREDICT) clinicaltrials.gov2019.
24. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195.
25. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction Models - Development, Evaluation, and Clinical Application. The New England journal of medicine 2020;382(17):1583-1586.
26. Zou Q, Xie S, Lin Z, Wu M, Ju Y. Finding the Best Classification Threshold in Imbalanced Classification. Big Data Research 2016;5:2-8.
27. Mateen BA, Liley J, Denniston AK, Holmes CC, Vollmer SJ. Improving the quality of machine learning in health applications and clinical research. Nature Machine Intelligence 2020;2(10):554-556.
28. Pedersen M, Verspoor K, Jenkinson M, Law M, Abbott DF, Jackson GD. Artificial intelligence for clinical decision support in neurology. Brain Communications 2020;2(2).
29. Shah NH, Milstein A, Bagley Ph DS. Making Machine Learning Models Clinically Useful. Jama 2019.
30. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. npj Digital Medicine 2018;1(1):40.
31. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.
32. Pietzsch JB, Garner AM, Marks WJ, Jr. Cost-Effectiveness of Deep Brain Stimulation for Advanced Parkinson's Disease in the United States. Neuromodulation 2016;19(7):689-697.
33. Lin HY, Hasegawa H, Mundil N, Samuel M, Ashkan K. Patients' Expectations and Satisfaction in Subthalamic Nucleus Deep Brain Stimulation for Parkinson Disease: 6-Year Follow-up. World neurosurgery 2019;121:e654-e660.
34. Kostoglou K, Michmizos KP, Stathis P, Sakas D, Nikita KS, Mitsis GD. Classification and Prediction of Clinical Improvement in Deep Brain Stimulation From Intraoperative Microelectrode Recordings. IEEE transactions on bio-medical engineering 2017;64(5):1123-1130.
35. Horvath K, Aschermann Z, Acs P, et al. Minimal clinically important difference on the Motor Examination part of MDS-UPDRS. Parkinsonism & related disorders 2015;21(12):1421-1426.
36. Schrag A, Sampaio C, Counsell N, Poewe W. Minimal clinically important change on the unified Parkinson's disease rating scale. Movement disorders : official journal of the Movement Disorder Society 2006;21(8):1200-1207.
37. Makkos A, Kovacs M, Aschermann Z, et al. Are the MDS-UPDRS-Based Composite Scores Clinically Applicable? Mov Disord 2018;33(5):835-839.
38. Shulman LM, Gruber-Baldini AL, Anderson KE, Fishman PS, Reich SG, Weiner WJ. The clinically important difference on the unified Parkinson's disease rating scale. Arch Neurol 2010;67(1):64-70.
39. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577-1579.
40. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC medicine 2015;13(1):1.

Part B:

Real-life monitoring of DBS motor response in Parkinson's disease

Chapter 4

An update on adaptive deep brain stimulation in Parkinson's disease

Habets JGV* & Heijmans M*, Kuijf M,
Janssen MLF, Temel Y, Kubben P
* Authors contributed equally

Published in Movement Disorders 2018, 33(12)
Doi: https://doi.org/10.1002/mds.115

Abstract

Advancing conventional open-loop DBS as a therapy for Parkinson's disease (PD) is crucial for overcoming important issues such as the delicate balance between beneficial and adverse effects and limited battery longevity that are currently associated with the treatment. Closed-loop or adaptive DBS (aDBS) aims to overcome these limitations by real-time adjustment of stimulation parameters based on continuous feedback input signals that are representative of the patient's clinical state. The focus of this update is to discuss the most recent developments regarding potential input signals and possible stimulation parameter modulation for aDBS in PD. Potential input signals for aDBS include basal ganglia local field potentials, cortical recordings (electrocorticography), wearable sensors, and eHealth and mHealth devices. Further, aDBS can be applied with different approaches of stimulation parameter modulation, the feasibility of which can be adapted depending on specific PD phenotypes. The implementation of technological developments like machine learning show potential in the design of such approaches, however energy consumption deserves further attention. Further, we discuss future considerations regarding the clinical implementation of aDBS in PD.

Introduction

Conventional deep brain stimulation (cDBS) of the subthalamic nucleus (STN) or the globus pallidus internus (GPi) is an established treatment for advanced stage Parkinson's disease (PD). While cDBS improves the motor symptoms of PD in both the short and long-term, it is not without limitations.[1, 2] Stimulation-induced side-effects such as dysarthria,[3] imbalance, and dyskinesia can occur and often require regular adjustments in stimulation, especially in the first phase after surgery.[4] Moreover, cDBS has limited battery life. These limitations have led to development and expanding scientific interest in closed-loop, responsive, or adaptive DBS (aDBS) (figure 1). For consistency reasons, only the term aDBS will be used.

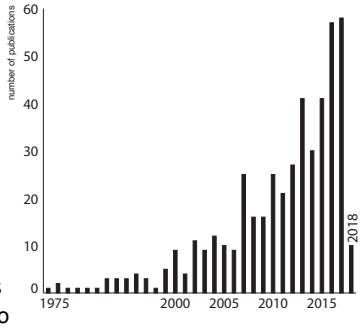


Figure 1: Yearly number of publications on adaptive deep brain stimulation in Parkinson's disease. Searched on Pubmed on 5-3-2018, using search command: [(parkinson*) AND (adaptive OR (closed loop) OR (closed-loop) OR responsive) AND (dbs OR stimulation)].

In cDBS, stimulation parameters are traditionally programmed and evaluated by a clinician during outpatient visits. If necessary, stimulation parameters are adjusted, and patients can perform minor changes within pre-set ranges themselves later. The goal of aDBS is to optimize this process further and automatically adapt stimulation parameters to the fluctuating clinical state of the patient, where in theory, stimulation is given only when necessary. As such, aDBS may generate fewer side-effects due to the possible decrease in energy given. In addition, although more power may be needed for data processing, the required battery consumption for stimulation potentially decreases and could result in increased battery longevity. Clinical proof-of-concept studies have already shown beneficial results using electrophysiological and/or wearable sensor recordings as feedback signals for aDBS in PD.[5, 6] The next step is to confirm whether such an approach continues its efficacy in the long term and to discuss new issues on the design of aDBS.

The development of a valid aDBS system in PD faces major challenges such as creating suitable input and processing input signals into beneficial output. In the following sections we present an update, future needs and possibilities for input signals, and stimulation paradigms for aDBS in PD. Much of the technological and clinical knowledge and experience discussed here also relates to the use of aDBS in other fluctuating neurologic and psychiatric diseases such as essential tremor (ET), dystonia, epilepsy, Tourette syndrome, and obsessive-compulsive disorder.

Potential input signals for aDBS

To develop a valid aDBS system, robust input signals representing the main PD symptoms are needed (figure 2). Symptoms vary from patient to patient, and therefore the suitability of input signals differs for individual patients (figure 3). Further, the necessity of supplementary (non-)invasive implants or devices, as well as additional processing and computational demands should be taken into consideration when comparing input signals. A comprehensive and concise overview of rationales and basic principles regarding potential input signals has recently been reviewed elsewhere.[7] We will elaborate further on previous work by discussing up-to-date progress and the remaining challenges regarding potential input signals for aDBS in PD.

aDBS based on electrophysiological recordings
Basal ganglia recordings
New generation DBS pulse generators can record local field potentials (LFPs), which have been correlated with clinical symptoms in several studies. For example, decreased beta band (8-35 Hz) activity in the STN by dopaminergic medication and/or DBS has been correlated with improved akinesia, bradykinesia, and rigidity,[8, 9] but not with tremor.[10-12] However, others found a correlation between STN-LFP recordings and tremor.[13, 14] Further, freezing-of-gait-periods [15] and differentiation between speech and movement activities [16] can be detected using STN-LFPs. Such differentiation of clinical indications underlines the potential of STN-LFP recordings as promising input signals, with the added benefit of not requiring additional implants or equipment compared to cDBS.[17]

A number of proof-of-concept studies using beta-LFPs to modify aDBS have shown motor improvement,[6, 18] less speech impairment,[19] and less levodopa-induced dyskinesia compared to cDBS,[20] which suggests this is a more efficient and effective method of stimulation. Moreover, a recent study demonstrated the feasibility and beneficial effects on motor symptoms of aDBS over the course of eight hours in akinetic-rigid PD patients.[21] Previous studies had already shown aDBS was applicable and effective in a freely moving[22] and a chronically implanted PD patient.[23]

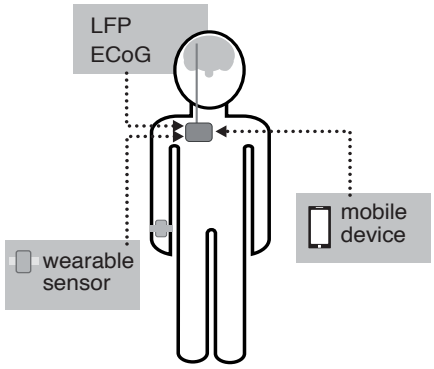


Figure 2: Schematic overview of the most used possible input signal origins for adaptive deep brain stimulation in PD. Sensors can also be worn on different locations, for example the chest, legs, or fingers. LFP: local field potential, recorded in the subthalamic nucleus. ECoG: electrocorticography.

Nevertheless, beta-LFPs in the STN are not (easily) detectable in all patients,[8] although this long-standing assumption has been contradicted recently.[24] Second, changes in beta-LFPs do not clearly capture *all* main symptoms of PD. For example, the relationship with tremor is debated. Although other STN-LFP signals like theta band (3-8 Hz) activity show promise in relation to tremor,[14] additional input signals to monitor tremor might be needed.[24, 25] Third, the clinical relevance or symptomatic contribution of high- versus low-beta-bands is a topic of discussion.[26, 27] Lastly, alpha/beta-band-activity is influenced by daily life events such as rest-tremor,[24] voluntary movements,[28, 29] movement artifacts during gait,[30] different vigilance states and sleep,[31] and aDBS itself,[32] which makes the isolation of disease-related signals difficult.

Despite these challenges, basal ganglia LFPs have been shown to function as a suitable input signal for aDBS. The main challenge to enable clinical use of LFPs is the development of standardized techniques that allow for automatic and validated interpretation of input signals. Therefore, further development of the hardware and software of aDBS systems is needed to acquire various frequency bands or additional input signals. Eventually, these sophisticated aDBS systems should

better suit the difference in clinical needs between akinetic-rigid and tremor-dominant PD patients.

Cortical recordings

A hallmark of PD is pathological hyperactivity of the corticobasal pathways, which is attributed to dopamine denervation of the striatum and substantia nigra.[33] This hyperactivity results in clinically identifiable cortical oscillations which can be measured invasively via electrocorticography (ECoG) using a subdural grid. aDBS can utilize these cortical oscillations as an input signal. For instance, one study showed that GPi-aDBS in non-human primates based on motor cortex (M1) beta-activity resulted in beneficial effects on akinesia.[34] Spatial-specific attenuation of cortical beta-hypersynchrony was also demonstrated in humans subsequent to STN-DBS.[35] Recent studies use phase-amplitude-coupling (PAC), whereby the amplitude of specific bandwidth-oscillations is coupled to specific oscillation-phases. In akinetic-rigid PD patients, excessive M1-beta-gamma-PAC decreased during STN-DBS, parallel to a decrease of clinically assessed bradykinesia.[36, 37] In contrast, in tremor-dominant PD patients, excessive M1-beta-PAC decreased during rest tremor.[38] Moreover, ECoG recordings showed potential to monitor dyskinesia,[39] gait-characteristics, such as walking duration and speed,[40] and to perform speech recognition.[41] The interpretation of cortical PAC-values therefore requires differentiation between phenotypic manifestations.

This work led to use of a fully implanted ECoG-based aDBS device in PD patients who experienced moderate dyskinesia despite optimized STN DBS therapy.[42] The authors adjusted stimulation voltage based on gamma-band (60 – 90 Hz) activity, which is related to dyskinesia. The clinical effect on bradykinesia and dyskinesia was maintained, while energy savings were ~40%.

A remaining concern is the limited correlation with PD symptoms and PAC attenuation due to movement preparation and execution.[37] Equal to beta-LFP in the STN, cortical beta-PAC is altered by DBS, which has implications for the analytic process.[43] Moreover, the time-frequency method used most often in PAC analysis might cause artifacts due to ignorance of the existence of both harmonic and non-sinusoidal neural dynamics in PD.[44] Another concern is that implantation of subdural grids may be associated with increased risk of complications such as hemorrhage and infection. As recently demonstrated, the use of cortical PAC is promising due to its potential ability to decode movement and behavior. Therefore, further steps are warranted to integrate the analyzed information from PAC and to develop analytic algorithms for different PD symptoms to perform aDBS based on cortical recordings in the whole PD spectrum.

Input signal vs. Parkinsonian symptoms	Correlation w/ tremor	Correlation w/ bradykinesia/ rigidity	Correlation w/ (freezing of) gait	Correlation w/ dyskinesia	Correlation w/ non-motor symptoms
Subcortical recordings (STN LFP)	-repeated reports -small evidence -on debate ●●●●	-repeated reports -reproduced evidence -starting consensus ●●●●●	-first reports -small evidence -no consensus ●●	-first reports -small evidence -no consensus ●●	-first reports -no evidence -no consensus ●●
Cortical recordings (ECoG)	-first reports -no evidence -no consensus ●●	-repeated reports -reproduced evidence -starting consensus ●●●●●	-not possible yet* ●	-first reports -small evidence ●●	-not possible yet * ●
Wearable sensors (accelerometer, gyroscope)	-repeated reports -reproduced evidence -starting consensus ●●●●●	-first reports -small evidence -no consensus ●●	-first reports -small evidence -no consensus ●●	-first reports -small evidence -no consensus ●●	-not possible yet ●
Mobile application	-first reports -no evidence -no consensus ●●	-repeated reports -no evidence -no consensus ●●●	-first reports -no evidence -no consensus ●●	-first reports -no evidence -no consensus ●●	-repeated reports -small evidence -no consensus ●●●

* In Parkinsonian patients, repeatedly reported with small evidence in other diseases.

Figure 3: Overview of published evidence of the feasibility of different input signals regarding different Parkinsonian symptoms for adaptive deep brain stimulation in Parkinson's disease. All input signals are scored on three categories per symptom. For each category 0, 0.5 or 1 bullet is given and the sum of them is visualized.
The first line indicates the amount of publications: not possible yet (0), first reports (0.5), repeated reports (1). The second line indicates the quality of reported evidence: no evidence (0), small evidence (0.5), reproduced evidence (1). The third line indicates the amount of consensus on the use of an input signal for a symptom: no consensus (0), on debate (0.5), starting consensus (1).

Surface electromyography

For several decades, surface electromyography (sEMG) signals have been used in tremor detection and more recently in tremor prediction.[45-48] Therefore, sEMG is considered to be a potential input signal for aDBS for ET and tremor-dominant PD. sEMG-based aDBS was feasible, effective, and efficient in ET patients.[49-51] Since the evidence of sEMG-based bradykinesia and rigidity detection methods is limited,[52, 53] sEMG should be combined with other input signals for akinetic-rigid PD patients. Another major concern of sEMG based aDBS is the potential loss of data quality due to the required self-management of sEMG sensors by patients. Furthermore, the signals must be processed and transmitted wirelessly to the pulse generator, which in turn may limit its battery life. To overcome these disadvantages, wireless sEMG sensors should be developed to withstand high contact impedances by using, for example, interchangeable patches to attach them to the skin, or subcutaneous implantable EMG electrodes. However, the limited potential of sEMG as an input signal and the current progress in wearable sensor development seem to make sEMG impractical for aDBS in PD.

aDBS based on neurochemical recordings

As stated in previous work, the development of aDBS based on neurochemical recordings is in an early phase.[7] Artifact-free neurochemical recordings were possible during DBS in rodents,[54] and dopamine fluctuations depending on DBS were found.[55] Therefore, neurochemical recordings were regarded to be potential input signal for aDBS, however the relationship between neurochemical recordings, PD symptoms, and DBS in humans has not been explored. Since no progress has been reported recently, the limitations for clinical use of neurochemical feedback in aDBS remain substantial.

aDBS based on wearable sensors

Monitoring PD symptoms through wearable sensors, or 'wearables', containing accelerometers and/or gyroscopes has gained considerable interest, and important progress has been made in the last decade.[56] Wearables are successful in predicting and detecting tremor[46, 48, 57] and show promise in assessing freezing-of-gait,[58] bradykinesia, and dyskinesia.[59-61]

Numerous studies based on tremor detection have supported the feasibility, effectiveness, and efficiency of wearables-based aDBS.[5, 51, 62] However, no other PD symptoms are yet detectable or implemented with wearable aDBS systems, and therefore the applicability for akinetic-rigid PD patients is unclear.[62]

The application of wearables for aDBS will rely heavily on machine learning approaches for distinguishing symptoms from voluntary movements.[63] Another concern is that patients will need to wear the sensors almost chronically. However, since sensors are getting smaller and more aesthetically attractive, this might not be a problem for all. In addition, continuous assessment of PD symptoms at home using wearables does not affect the health-related quality of life.[64] Lastly, signal processing and wireless data transmission may limit battery life of wearables and pulse generators.

To implement aDBS controlled by wearables, algorithms to monitor other cardinal motor-symptoms than tremor need further development and clinical validation to expand the potential for akinetic-rigid PD patients. For tremor-dominant patients, clinical trials with longer follow-up periods should be done to prove superiority compared to cDBS.

aDBS based on PD monitoring systems including eHealth and mHealth applications
Wearables and electrophysiological recordings disregard the subjective experience of motor symptoms and the assessment of non-motor symptoms. We believe subjective experience of motor symptoms could improve the interpretation of objective motor symptom monitoring. Non-motor symptoms are important for quality of life scores and might predict overall DBS outcomes.[65] Electronic health (eHealth) and mobile health (mHealth) applications and telemonitoring concepts have been recently integrated into PD care and contain the aforementioned missing features.[66-69] Most of these developments are achieved in order to improve PD care and to ensure its accessibility and cost-effectiveness.[70, 71] However, these developments also hold promise for aDBS.

A recent trial demonstrated that cDBS-setting adjustment via telemonitoring was feasible.[72] Introducing automated monitoring and increasing the frequency of DBS-setting adjustment brings this concept close to (semi-continuous) aDBS. The lack of valid continuous PD monitoring tools led to development of multimodal PD monitoring systems. These systems include, for example, wearables and mobile applications and distinguish themselves from systems discussed above by adding assessments of cognition, speech, subjective disease burden, and active motor tasks. This potential was recently underlined by development of a smartphone application to capture symptom fluctuation during the day.[73]

Several recently initiated trials test the feasibility and clinical value of multimodal PD monitoring systems in the patient's home environment. So far, these systems aim to differentiate ON/OFF states via wearables and a diary,[74] detect the need for changes in or improve adherence of pharmacological therapy,[75-77] and monitor clinical well-being in a holistic fashion.[78, 79] Other systems aim to detect relevant neurophysiological biomarkers for home monitoring in order to improve postoperative DBS care[67] and to assess the effect of DBS parameter adjustments with wearables.[69] Further, the feasibility of the experience sampling method is demonstrated among

PD patients.[80] This method collects subjective experiences of both motor and non-motor symptoms multiple times a day during the flow of daily life.

The abovementioned studies show the feasibility of using multimodal monitoring systems among PD patients. We believe there might be a role for such multimodal PD monitoring systems in aDBS, because they have the potential to combine subjective assessments of burden and non-motor symptoms with objective input signals. Especially during the initial post-operative phase, combining these input signals may be of great value for adjusting DBS. The feasibility of such a holistic approach should be explored further.

Stimulation parameter modulation in aDBS

The process of collecting continuous data representing (non-)motor symptoms is the first major challenge for developing an aDBS system for PD. A second major challenge is the design of a system which automates the complex reasoning and decision making currently achieved by clinicians, which will require more advanced and distinctive signal processing than is currently available. This challenge contains several issues, such as the frequency of stimulation parameter adjustments, the nature of stimulation parameter adjustments, data transfer, data computation, and battery consumption.

Most aDBS research has so far focused on potential input signals, and therefore several issues regarding the design of stimulation parameter modulation are less well studied. In the following sections, we discuss the current progress and remaining challenges regarding stimulation parameter modulation in aDBS.

Amplitude modulation approaches

All reported aDBS systems in PD until now are based on automatic amplitude modulation (AM). AM can be applied in different designs. *ON/OFF AM* is an aDBS paradigm which varies between periods during which stimulation is given with a predefined amplitude and a set frequency and pulse width, and periods during which stimulation is switched off (figure 4A). ON/OFF AM systems studied in akinetic-rigid PD patients applied stimulation as long as the beta-LFP recorded in the STN exceeded a certain threshold.[6, 28] In contrast, ON/OFF AM systems studied in tremor-dominant PD patients were designed to start a stimulation-period several seconds before tremor re-occurrs based on tremor prediction using machine learning algorithms.[46, 48] Since tremor should not reoccur during stimulation, there is no feedback signal which identifies the end of the stimulation-period. Recent research has shown a stimulation duration of 30 seconds led to a ratio of stimulation time versus tremor-free time off-stimulation over 50% in one-third of patients.[62] Future research will need to clarify how ON/OFF AM can be implemented optimally for different PD phenotypes and different input signals.

When using ON/OFF AM, other details should be considered. First, a ramping onset, which increases the stimulation voltage from zero toward a predefined amplitude can be used to overcome paresthesia.[6, 32] Further, ON/OFF AM can be applied in a phase-dependent manner, in which a stimulus is applied with a fixed latency to an input signal.[25] Phase-dependent aDBS is hypothesized to have advantages over standard aDBS. Increased clinical benefit is suggested by targeting specific pathological neurophysiological phases in PD.[6, 32, 34] Also, phase-dependent aDBS might induce long-lasting beneficial effects due to possible long-term potentiation/depotentiation in the STN.[81] Moreover, phase-dependent aDBS reduced tremor severity and prevented breakthrough tremor while consuming less energy compared to cDBS.[82] Studies assessing these suggested advantages of phase-dependency in aDBS are required.

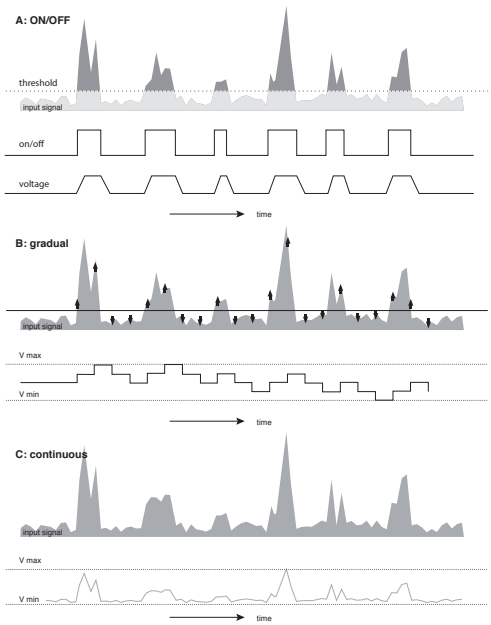


Figure 4: Schematic overview of different amplitude modulation paradigms used in adaptive deep brain stimulation in Parkinson's disease. A: ON/OFF paradigm, which stimulates with ramping onset when input signals exceed a certain threshold. B: Gradual paradigm, which increases or decreases stimulation amplitude stepwise when input signal exceeds or does not exceed a certain threshold respectively. C: Continuous paradigm, which modifies stimulation amplitude according to strength of input signal.

Other AM aDBS paradigms use a gradual or a continuous AM approach. *Gradual AM* increases or decreases the amplitude stepwise when the input signal is respectively higher or lower than certain thresholds (figure 4B).[5, 42] Minimal and maximal stimulation amplitudes and the voltage change per step have yet to be defined. Recently, a gradual AM approach based on tremor power introduced two feedback loop computations. One slow-loop gradually adjusted the amplitude-baseline to prevent re-emergence of diminished tremor, and one fast-loop adjusted the actual amplitude rapidly to mitigate occurring tremor.[83] This design will need to be reproduced, and the added benefit should be assessed. *Continuous AM* links every possible input signal to a corresponding preset output amplitude (fig. 4C). Thus, the output amplitude inclines toward a parallel line of the input signal.[21, 22]

It is clear that stimulation parameter modulation can be applied in several ways in aDBS. At this moment, no research has been done to compare different approaches in general, or for specific phenotypes or input signals. In the next paragraph, we will elaborate on the clinical demands towards stimulation parameter modulation in aDBS per phenotype.

Stimulation parameter modulation demands per phenotype

An optimally performing aDBS system should prevent overstimulation during periods with less symptoms, and it should increase voltage in a timely manner to minimize the duration and severity of symptomatic periods. Therefore, the frequency of input signal evaluation can be an important difference in stimulation parameter modulation according to phenotype. This should be based on the frequency at which the monitored symptom is expected to fluctuate or reoccur after stopping or decreasing stimulation.

In tremor-dominant PD, an aDBS system should ideally stimulate on 'tremor-control' level before the tremor actually occurs. Since tremor fluctuates rapidly, the AM approach should rapidly respond to tremor reoccurrence in order to minimize tremor duration.
Compared to ON/OFF AM, a gradual or continuous AM approach needs more evaluation before stimulation reaches 'tremor-control' level. However, ON/OFF AM always stimulates with the preset voltage, which might cause overstimulation.
In developing an optimal AM approach, the development of tremor prediction machine learning models is important.[48] At the moment, most of these models are very accurate in scaling the tremor severity rather than predicting the reoccurrence.[84] Future studies should analyze different evaluation frequencies, corresponding computational costs, and tremor reduction to compare the feasibility of different AM approaches.

In akinetic-rigid PD, motor symptom fluctuations will be less frequent and less acute. Different input signal evaluation frequencies in aDBS for akinetic-rigid patients have not yet been compared. Whether gradual or continuous, AM is superior to ON/OFF AM in this group. It is dependent on improved symptom reduction and prevention of over-stimulation when stimulating between zero and maximal amplitude.
Recent work on the modulatory effect of aDBS on beta-LFP suggests ON/OFF AM to be better suited for akinetic-rigid patients than gradual AM.[32] They found a correlation between longer beta-bursts (>0.6 seconds) and clinical impairment. Consequently, this implies that these longer beta bursts should trigger stimulation, and rapid anticipation and frequent evaluation of beta power is thus needed. Due to this required rapid anticipation, they prefer ON/OFF AM.
However, if the input signal follows the rhythm of akinesia and rigidity fluctuations, the input signal evaluation frequency could decrease, and a gradual AM approach might also be suitable and efficient. Whether this less frequent evaluation is feasible with STN-LFP recordings has not been explored. At the moment, wearable sensors might have more potential to accomplish this compared to STN-LFP recordings.
In aDBS for PD patients suffering moderate dyskinesia, these considerations were also addressed.[42] The authors saw aDBS transitions more frequently than expected based on clinical symptomatology. They suggested a slow ramping onset of stimulation voltage adjustments or alternative use of the triggering threshold, for example a higher threshold or a two-step threshold, in order to prevent too frequent stimulation parameter adjustments.

Future considerations

Issues for clinical implementation

As discussed above, the research field on aDBS in PD is rapidly evolving (figure 1). In this section, we will highlight additional prospective issues that should be solved to realize a feasible aDBS system for chronic therapy.

Individual expectations and desires regarding an aDBS system can differ due to inter-individual differences in the clinical course and personal coping strategies in PD patients. Individually tailored aDBS paradigms should respond to these factors, particularly aDBS systems that enable personal nuances in stimulation parameter modulation. However, aDBS in PD first needs a feasible standard system and stimulation parameter modulation, or one aDBS system per phenotype, before individualized fine-tuning can take place. It is plausible that each individual will start aDBS therapy with a calibration period, similar to cDBS therapy. Instead of a trial and error period trying different amplitudes, frequencies, or electrode-contacts, the aDBS calibration period might try out different frequencies of stimulation parameter modulation, different threshold levels, or different voltage-steps per modulation. Ideally, this process is automated by a self-regulating algorithm.

aDBS during sleep

Although not discussed yet, a feasible aDBS system should consider the differences in patient preference and input signal during sleep. Akinetic-rigid patients might consider stimulation at night as important, since they suffer from rigidity at night and in the morning. Tremor-dominant patients might need less stimulation at night due to a lower disease burden.

Regarding input signals, electrophysiological signals are influenced by vigilance state and therefore deserve different interpretation during sleep periods than during awake periods.[31] Also, wearable sensors might be programmed with 'sleep' or 'rest' detection algorithms which initiates a specific 'sleep-stimulation paradigm'.

Monitoring of non-motor symptoms and side-effects

In general, current aDBS input signals are focused on motor symptom detection to evaluate the therapeutic effect. As discussed before, a first step towards personalized therapy can be to develop different aDBS approaches for the different main motor symptoms per phenotype. Future designs might expand the specificity per phenotype by considering non-motor symptoms[85] and potentially side effects caused by aDBS, like autonomic functions, dyskinesia, or speech deterioration. Including these features will make data analysis even more complicated. This future challenge requires complicated nuances which are out of reach for aDBS presently.

aDBS modulation other than amplitude modulation

Later aDBS systems might explore the use of different stimulation parameter modulations for specific clinical situations, e.g. frequency modulation (FM). Possibly, stimulation parameters in bilateral aDBS could be evaluated and adjusted per side separately, tailoring aDBS per side.

The application of FM is hypothesized to contribute to tailored DBS paradigms.[86] Three recent reviews on low-frequency STN-DBS described beneficial effects on freezing-of-gait, speech, and swallowing that did not respond to, or were caused by, high frequency DBS. However, beneficial effects could not always be reproduced, and low frequency stimulation sometimes led to worsening of cardinal PD symptoms.[87-89] The effect of variable frequency stimulation, a paradigm interleaving high and low frequency DBS,[90] will be explored soon.[91]

Also, pulse width modulation might provide clinical benefit in certain situations. By exciting thin axon bundles belonging to the direct cortico-subthalamic pathway more selectively,[92] therapeutic windows may increase using shorter pulse-widths, while using less energy.[93, 94]

Battery power balance

aDBS may require less battery power for stimulation compared to cDBS. In contrast, more battery power may be needed for data processing and transferal, for example, by Bluetooth®. There are several options to minimize the additional power needed by the pulse generator and to eventually make battery replacement less frequent. First, comparing the computational demands of various signal processing and machine learning approaches should minimize the required power.[16] Second, the possibility to perform analyses on external devices or cloud-platform solutions should be evaluated. The energy saved by outsourcing these computations should be compared with the energy required of wireless data transfer. Third, rechargeable pulse generators should be further developed regarding clinical applicability.[95]

Socio-economical relevance

We suggest that implementation of aDBS systems in PD care will result in beneficial socio-economic effects. Most importantly, if aDBS results in an improved ratio between beneficial and side effects, quality of life will improve and patients will function better in society. Also, the economic burden will decrease, since PD patients can be part of the working population for a longer period and need less care.

Conclusion

Although impressive progress in aDBS for PD has been made over the last decade, major challenges to chronic application are still pending. We believe research into clinical associations of input signals should concentrate on different PD phenotypes. Since the correlation of different input signals with PD symptomatology varies (fig. 3), we believe no single currently available input signal will cover the heterogeneity of all phenotypes in PD patients. To achieve this ambition, thoughtful combining and selection of input signals is inevitable. The increasing trend of combining knowledge between neurologists, neurosurgeons, engineers, and computer scientists is crucial in this field and opens the gate to translational medicine 2.0: “from byte to bedside”.

Acknowledgment

No acknowledgments

References

1. Janssen ML, Duits AA, Turaihi AH, et al. Subthalamic nucleus high-frequency stimulation for advanced Parkinson's disease: motor and neuropsychological outcome after 10 years. Stereotactic and functional neurosurgery 2014;92(6):381-387.
2. Limousin P, Pollak P, Benazzouz A, et al. Effect of parkinsonian signs and symptoms of bilateral subthalamic nucleus stimulation. Lancet 1995;345(8942):91-95.
3. Tripoliti E, Zrinzo L, Martinez-Torres I, et al. Effects of subthalamic stimulation on speech of consecutive patients with Parkinson disease. Neurology 2011;76(1):80-86.
4. Deuschl G, Schade-Brittinger C, Krack P, et al. A randomized trial of deep-brain stimulation for Parkinson's disease. The New England journal of medicine 2006;355(9):896-908.
5. Malekmohammadi M, Herron J, Velisar A, et al. Kinematic Adaptive Deep Brain Stimulation for Resting Tremor in Parkinson's Disease. Movement disorders : official journal of the Movement Disorder Society 2016;31(3):426-428.
6. Little S, Pogosyan A, Neal S, et al. Adaptive deep brain stimulation in advanced Parkinson disease. Annals of neurology 2013;74(3):449-457.
7. Arlotti M, Rosa M, Marceglia S, Barbieri S, Priori A. The adaptive deep brain stimulation challenge. Parkinsonism & related disorders 2016;28:12-17.
8. Giannicola G, Marceglia S, Rossi L, et al. The effects of levodopa and ongoing deep brain stimulation on subthalamic beta oscillations in Parkinson's disease. Experimental neurology 2010;226(1):120-127.
9. Little S, Pogosyan A, Kuhn AA, Brown P. beta band stability over time correlates with Parkinsonian rigidity and bradykinesia. Experimental neurology 2012;236(2):383-388.
10. Kuhn AA, Kupsch A, Schneider GH, Brown P. Reduction in subthalamic 8-35 Hz oscillatory activity correlates with clinical improvement in Parkinson's disease. Eur J Neurosci 2006;23(7):1956-1960.
11. Chen CC, Hsu YT, Chan HL, et al. Complexity of subthalamic 13-35 Hz oscillatory activity directly correlates with clinical impairment in patients with Parkinson's disease. Experimental neurology 2010;224(1):234-240.
12. Beudel M, Oswal A, Jha A, et al. Oscillatory Beta Power Correlates With Akinesia-Rigidity in the Parkinsonian Subthalamic Nucleus. Movement disorders : official journal of the Movement Disorder Society 2017;32(1):174-175.
13. Hirschmann J, Schoffelen JM, Schnitzler A, van Gerven MAJ. Parkinsonian rest tremor can be detected accurately based on neuronal oscillations recorded from the subthalamic nucleus. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 2017;128(10):2029-2036.
14. Contarino MF, Bour LJ, Bot M, et al. Tremor-specific neuronal oscillation pattern in dorsal subthalamic nucleus of parkinsonian patients. Brain stimulation 2012;5(3):305-314.
15. Syrkin-Nikolau J, Koop MM, Prieto T, et al. Subthalamic neural entropy is a feature of freezing of gait in freely moving people with Parkinson's disease. Neurobiology of Disease 2017;108(Supplement C):288-297.
16. Golshan HM, Hebb AO, Hanrahan SJ, Nedrud J, Mahoor MH. A hierarchical structure for human behavior classification using STN local field potentials. J Neurosci Methods 2018;293:254-263.
17. Neumann WJ, Staub F, Horn A, et al. Deep Brain Recordings Using an Implanted Pulse Generator in Parkinson's Disease. Neuromodulation : journal of the International Neuromodulation Society 2016;19(1):20-24.
18. Little S, Beudel M, Zrinzo L, et al. Bilateral adaptive deep brain stimulation is effective in Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 2016;87(7):717-721.
19. Little S, Tripoliti E, Beudel M, et al. Adaptive deep brain stimulation for Parkinson's disease demonstrates reduced speech side effects compared to conventional stimulation in the acute setting. Journal of Neurology, Neurosurgery & Psychiatry 2016;87(12):1388-1389.
20. Rosa M, Arlotti M, Marceglia S, et al. Adaptive deep brain stimulation controls levodopa-induced side effects in Parkinsonian patients. Movement disorders : official journal of the Movement Disorder Society 2017;32(4):628-629.
21. Arlotti M, Marceglia S, Foffani G, et al. Eight-hours adaptive deep brain stimulation in patients with Parkinson disease. Neurology 2018.
22. Rosa M, Arlotti M, Ardolino G, et al. Adaptive deep brain stimulation in a freely moving Parkinsonian patient. Movement disorders : official journal of the Movement Disorder Society 2015;30(7):1003-1005.
23. Pina-Fuentes D, Little S, Oterdoom M, et al. Adaptive DBS in a Parkinson's patient with chronically implanted DBS: A proof of principle. Movement disorders : official journal of the Movement Disorder Society 2017;32(8):1253-1254.
24. Shreve LA, Velisar A, Malekmohammadi M, et al. Subthalamic oscillations and phase amplitude coupling are greater in the more affected hemisphere in Parkinson's disease. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 2017;128(1):128-137.
25. Meidahl AC, Tinkhauser G, Herz DM, Cagnan H, Debarros J, Brown P. Adaptive Deep Brain Stimulation for Movement Disorders: The Long Road to Clinical Therapy. Movement Disorders 2017;32(6):810-819.

26. van Wijk BC, Beudel M, Jha A, et al. Subthalamic nucleus phase-amplitude coupling correlates with motor impairment in Parkinson's disease. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 2016;127(4):2010-2019.
27. Blumenfeld Z, Koop MM, Prieto TE, et al. Sixty-hertz stimulation improves bradykinesia and amplifies subthalamic low-frequency oscillations. Movement disorders : official journal of the Movement Disorder Society 2017;32(1):80-88.
28. Johnson LA, Nebeck SD, Muralidharan A, Johnson MD, Baker KB, Vitek JL. Closed-Loop Deep Brain Stimulation Effects on Parkinsonian Motor Symptoms in a Non-Human Primate - Is Beta Enough? Brain stimulation 2016;9(6):892-896.
29. Storzer L, Butz M, Hirschmann J, et al. Bicycling suppresses abnormal beta synchrony in the Parkinsonian basal ganglia. Annals of neurology 2017;82(4):592-601.
30. Hell F, Taylor PCJ, Mehrkens JH, Botzel K. Subthalamic stimulation, oscillatory activity and connectivity reveal functional role of STN and network mechanisms during decision making under conflict. NeuroImage 2018;171:222-233.
31. Escobar D, Johnson LA, Nebeck SD, et al. Parkinsonism and Vigilance: Alteration in neural oscillatory activity and phase-amplitude coupling in the basal ganglia and motor cortex. J Neurophysiol 2017:jn 00388 02017.
32. Tinkhauser G, Pogosyan A, Little S, et al. The modulatory effect of adaptive deep brain stimulation on beta bursts in Parkinson's disease. Brain : a journal of neurology 2017;140(4):1053-1067.
33. Hammond C, Bergman H, Brown P. Pathological synchronization in Parkinson's disease: networks, models and treatments. Trends Neurosci 2007;30(7):357-364.
34. Rosin B, Slovik M, Mitelman R, et al. Closed-loop deep brain stimulation is superior in ameliorating parkinsonism. Neuron 2011;72(2):370-384.
35. Whitmer D, de Solages C, Hill B, Yu H, Henderson JM, Bronte-Stewart H. High frequency deep brain stimulation attenuates subthalamic and cortical rhythms in Parkinson's disease. Frontiers in human neuroscience 2012;6:155.
36. de Hemptinne C, Ryapolova-Webb ES, Air EL, et al. Exaggerated phase-amplitude coupling in the primary motor cortex in Parkinson disease. Proceedings of the National Academy of Sciences of the United States of America 2013;110(12):4780-4785.
37. de Hemptinne C, Swann NC, Ostrem JL, et al. Therapeutic deep brain stimulation reduces cortical phase-amplitude coupling in Parkinson's disease. Nature neuroscience 2015;18(5):779-786.
38. Qasim SE, de Hemptinne C, Swann NC, Miocinovic S, Ostrem JL, Starr PA. Electrocorticography reveals beta desynchronization in the basal ganglia-cortical loop during rest tremor in Parkinson's disease. Neurobiol Dis 2016;86:177-186.
39. Swann NC, de Hemptinne C, Miocinovic S, et al. Gamma Oscillations in the Hyperkinetic State Detected with Chronic Human Brain Recordings in Parkinson's Disease. The Journal of neuroscience : the official journal of the Society for Neuroscience 2016;36(24):6445-6458.
40. McCrimmon CM, Wang PT, Heydari P, et al. Electrocorticographic Encoding of Human Gait in the Leg Primary Motor Cortex. Cereb Cortex 2017:1-11.
41. Herff C, Schultz T. Automatic Speech Recognition from Neural Signals: A Focused Review. Frontiers in neuroscience 2016;10:429.
42. Swann NC, de Hemptinne C, Thompson MC, et al. Adaptive deep brain stimulation for Parkinson's disease using motor cortex sensing. Journal of neural engineering 2018;15(4):046006.
43. Cole SR, van der Meij R, Peterson EJ, de Hemptinne C, Starr PA, Voytek B. Nonsinusoidal Beta Oscillations Reflect Cortical Pathophysiology in Parkinson's Disease. The Journal of neuroscience : the official journal of the Society for Neuroscience 2017;37(18):4830-4840.
44. Pittman-Polletta B, Hsieh WH, Kaur S, Lo MT, Hu K. Detecting phase-amplitude coupling with high frequency resolution using adaptive decompositions. J Neurosci Methods 2014;226:15-32.
45. Scholz E, Diener HC, Noth J, Friedemann H, Dichgans J, Bacher M. Medium and long latency EMG responses in leg muscles: Parkinson's disease. Journal of neurology, neurosurgery, and psychiatry 1987;50(1):66-70.
46. Basu I, Graupe D, Tuninetti D, et al. Pathological tremor prediction using surface electromyogram and acceleration: potential use in 'ON-OFF' demand driven deep brain stimulator design. Journal of neural engineering 2013;10(3):036019.
47. Camara C, Warwick K, Bruna R, Aziz T, del Pozo F, Maestu F. A Fuzzy Inference System for Closed-Loop Deep Brain Stimulation in Parkinson's Disease. Journal of medical systems 2015;39(11):155.
48. Khobragade N, Graupe D, Tuninetti D. Towards fully automated closed-loop Deep Brain Stimulation in Parkinson's disease patients: A LAMSTAR-based tremor predictor. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2015;2015:2616-2619.
49. Graupe D, Basu I, Tuninetti D, Vannemreddy P, Slavin KV. Adaptively controlling deep brain stimulation in essential tremor patient via surface electromyography. Neurol Res 2010;32(9):899-904.
50. Yamamoto T, Katayama Y, Ushiba J, et al. On-demand control system for deep brain stimulation for treatment of intention tremor. Neuromodulation : journal of the International Neuromodulation Society 2013;16(3):230-235; discussion 235.

51. Herron JA, Thompson MC, Brown T, Chizeck HJ, Ojemann JG, Ko AL. Chronic electrocorticography for sensing movement intention and closed-loop deep brain stimulation with wearable sensors in an essential tremor patient. Journal of neurosurgery 2017;127(3):580-587.
52. Askari S, Zhang M, Won DS. An EMG-based system for continuous monitoring of clinical efficacy of Parkinson's disease treatments. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2010;2010:98-101.
53. Levin J, Krafczyk S, Valkovic P, Eggert T, Claassen J, Botzel K. Objective measurement of muscle rigidity in Parkinsonian patients treated with subthalamic stimulation. Movement disorders : official journal of the Movement Disorder Society 2009;24(1):57-63.
54. Chang SY, Kimble CJ, Kim I, et al. Development of the Mayo Investigational Neuromodulation Control System: toward a closed-loop electrochemical feedback system for deep brain stimulation. Journal of neurosurgery 2013;119(6):1556-1565.
55. Grahn PJ, Mallory GW, Khurram OU, et al. A neurochemical closed-loop controller for deep brain stimulation: toward individualized smart neuromodulation therapies. Frontiers in neuroscience 2014;8:169.
56. Sanchez-Ferro A, Elshehabi M, Godinho C, et al. New methods for the assessment of Parkinson's disease (2005 to 2015): A systematic review. Movement disorders : official journal of the Movement Disorder Society 2016;31(9):1283-1292.
57. Delrobaei M, Memar S, Pieterman M, Stratton TW, McIsaac K, Jog M. Towards remote monitoring of Parkinson's disease tremor using wearable motion capture systems. J Neurol Sci 2018;384:38-45.
58. Rodriguez-Martin D, Sama A, Perez-Lopez C, et al. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. PloS one 2017;12(2):e0171764.
59. Griffiths RI, Kotschet K, Arfon S, et al. Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. Journal of Parkinson's disease 2012;2(1):47-55.
60. Delrobaei M, Baktash N, Gilmore G, McIsaac K, Jog M. Using Wearable Technology to Generate Objective Parkinson's Disease Dyskinesia Severity Score: Possibilities for Home Monitoring. IEEE Trans Neural Syst Rehabil Eng 2017;25(10):1853-1863.
61. Hasan H, Athauda DS, Foltynie T, Noyce AJ. Technologies Assessing Limb Bradykinesia in Parkinson's Disease. Journal of Parkinson's disease 2017;7(1):65-77.
62. Graupe D, Khobragade N, Tuninetti D, Basu I, Slavin KV, Verhagen Metman L. Who May Benefit From On-Demand Control of Deep Brain Stimulation? Noninvasive Evaluation of Parkinson Patients. Neuromodulation : journal of the International Neuromodulation Society 2018.
63. Kubota KJ, Chen JA, Little MA. Machine learning for large-scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. Movement disorders : official journal of the Movement Disorder Society 2016;31(9):1314-1326.
64. van Uem JM, Isaacs T, Lewin A, et al. A Viewpoint on Wearable Technology-Enabled Measurement of Wellbeing and Health-Related Quality of Life in Parkinson's Disease. Journal of Parkinson's disease 2016;6(2):279-287.
65. Dafsari HS, Weiss L, Silverdale M, et al. Short-term quality of life after subthalamic stimulation depends on non-motor symptoms in Parkinson's disease. Brain stimulation 2018.
66. Eberle W, Penders J, Yazicioglu RF. Closing the loop for Deep Brain Stimulation implants enables personalized healthcare for Parkinson's disease patients. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2011;2011:1556-1558.
67. Marceglia S, Rossi E, Rosa M, et al. Web-based telemonitoring and delivery of caregiver support for patients with Parkinson disease after deep brain stimulation: protocol. JMIR Res Protoc 2015;4(1):e30.
68. Zhang C, Li D, Zeljic K, Tan H, Ning Y, Sun B. A Remote and Wireless Deep Brain Stimulation Programming System. Neuromodulation : journal of the International Neuromodulation Society 2016;19(4):437-439.
69. Angeles P, Tai Y, Pavese N, Wilson S, Vaidyanathan R. Automated assessment of symptom severity changes during deep brain stimulation (DBS) therapy for Parkinson's disease. IEEE Int Conf Rehabil Robot 2017;2017:1512-1517.
70. 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(9):1258-1262.
71. Schneider RB, Biglan KM. The promise of telemedicine for chronic neurological disorders: the example of Parkinson's disease. The Lancet Neurology 2017;16(7):541-551.
72. Li D, Zhang C, Gault J, et al. Remotely Programmed Deep Brain Stimulation of the Bilateral Subthalamic Nucleus for the Treatment of Primary Parkinson Disease: A Randomized Controlled Trial Investigating the Safety and Efficacy of a Novel Deep Brain Stimulation System. Stereotactic and functional neurosurgery 2017;95(3):174-182.
73. 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.

74. Bayes A, Sama A, Prats A, et al. A "HOLTER" for Parkinson's disease: Validation of the ability to detect on-off states using the REMPARK system. Gait Posture 2017;59:1-6.
75. Lakshminarayana R, Wang D, Burn D, et al. Smartphone- and internet-assisted self-management and adherence tools to manage Parkinson's disease (SMART-PD): study protocol for a randomised controlled trial (v7; 15 August 2014). Trials 2014;15:374.
76. Tzallas AT, Tsipouras MG, Rigas G, et al. PERFORM: a system for monitoring, assessment and management of patients with Parkinson's disease. Sensors (Basel) 2014;14(11):21329-21357.
77. Ferreira JJ, Godinho C, Santos AT, et al. Quantitative home-based assessment of Parkinson's symptoms: the SENSE-PARK feasibility and usability study. BMC Neurol 2015;15:89.
78. Cancela J, Villanueva Mascato S, Gatsios D, et al. Monitoring of motor and non-motor symptoms of Parkinson's disease through a mHealth platform. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2016;2016:663-666.
79. Neurotechnologies GL. Kinesia ProViewTM. http://glneurotech.com/kinesia/products/proview/; 2017.
80. Broen MP, Marsman VA, Kuijf ML, Van Oostenbrugge RJ, van Os J, Leentjens AF. Unraveling the Relationship between Motor Symptoms, Affective States and Contextual Factors in Parkinson's Disease: A Feasibility Study of the Experience Sampling Method. PloS one 2016;11(3):e0151195.
81. Azodi-Avval R, Gharabaghi A. Phase-dependent modulation as a novel approach for therapeutic brain stimulation. Frontiers in computational neuroscience 2015;9:26.
82. Cagnan H, Pedrosa D, Little S, et al. Stimulating at the right time: phase-specific deep brain stimulation. Brain : a journal of neurology 2017;140(1):132-145.
83. Herron J, Velisar A, Malekmohammadi M, Bronte-Stewart H, Chizeck HJ. A Metric for Evaluating and Comparing Closed-Loop Deep Brain Stimulation Algorithms. arXiv preprint arXiv:160509312 2016.
84. Jeon H, Lee W, Park H, et al. High-accuracy automatic classification of Parkinsonian tremor severity using machine learning method. Physiol Meas 2017;38(11):1980-1999.
85. Marras C, Chaudhuri KR. Nonmotor features of Parkinson's disease subtypes. Movement disorders : official journal of the Movement Disorder Society 2016;31(8):1095-1102.
86. Fasano A, Lozano AM. The FM/AM world is shaping the future of deep brain stimulation. Movement disorders : official journal of the Movement Disorder Society 2014;29(2):161-163.
87. Baizabal-Carvallo JF, Alonso-Juarez M. Low-frequency deep brain stimulation for movement disorders. Parkinsonism & related disorders 2016;31(Supplement C):14-22.
88. Dayal V, Limousin P, Foltynie T. Subthalamic Nucleus Deep Brain Stimulation in Parkinson's Disease: The Effect of Varying Stimulation Parameters. Journal of Parkinson's disease 2017;7(2):235-245.
89. Xie T, Padmanaban M, Bloom L, et al. Effect of low versus high frequency stimulation on freezing of gait and other axial symptoms in Parkinson patients with bilateral STN DBS: a mini-review. Translational neurodegeneration 2017;6:13.
90. Jia F, Guo Y, Wan S, et al. Variable frequency stimulation of subthalamic nucleus for freezing of gait in Parkinson's disease. Parkinsonism & related disorders 2015;21(12):1471-1472.
91. Jia F, Hu W, Zhang J, et al. Variable frequency stimulation of subthalamic nucleus in Parkinson's disease: Rationale and hypothesis. Parkinsonism & related disorders 2017;39(Supplement C):27-30.
92. Reich MM, Steigerwald F, Sawalhe AD, et al. Short pulse width widens the therapeutic window of subthalamic neurostimulation. Ann Clin Transl Neurol 2015;2(4):427-432.
93. Akbar U, Raike RS, Hack N, et al. Randomized, Blinded Pilot Testing of Nonconventional Stimulation Patterns and Shapes in Parkinson's Disease and Essential Tremor: Evidence for Further Evaluating Narrow and Biphasic Pulses. Neuromodulation : journal of the International Neuromodulation Society 2016;19(4):343-356.
94. Steigerwald F, Timmermann L, Kuhn A, et al. Pulse duration settings in subthalamic stimulation for Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2018;33(1):165-169.
95. Rizzi M, Messina G, Penner F, D'Ammando A, Muratorio F, Franzini A. Internal Pulse Generators in Deep Brain Stimulation: Rechargeable or Not? World Neurosurg 2015;84(4):1020-1029.

Chapter 5

Monitoring Parkinson's disease symptoms during daily life: a feasibility study

Heijmans M, Habets JGV, Herff C,
Aarts J, Stevens A, Kuijf ML, Kubben PL

Published in npj Parkinson's Disease 2019, 5(21)
DOI: https://doi.org/10.1038/s41531-019-0093-5

Abstract

Parkinson's disease symptoms are most often charted using the MDS-UPDRS. Limitations of this approach include the subjective character of the assessments and a discrepant performance in the clinic compared to the home situation. Continuous monitoring using wearable devices is believed to eventually replace this golden standard, but measurements often lack a parallel ground truth or are only tested in lab settings. To overcome these limitations, this study explores the feasibility of a newly developed Parkinson's disease monitoring system, which aims to measure Parkinson's disease symptoms during daily life by combining wearable sensors with an experience sampling method application. Twenty patients with idiopathic Parkinson's disease participated in this study. During a period of two consecutive weeks, participants had to wear three wearable sensors and had to complete questionnaires at seven semi-random moments per day on their mobile phone. Wearable sensors collected objective movement data, and the questionnaires containing questions about amongst others Parkinson's disease symptoms served as parallel ground truth. Results showed that participants wore the wearable sensors during 94% of the instructed timeframe and even beyond. Furthermore, questionnaire completion rates were high (79,1%) and participants evaluated the monitoring system positively. A preliminary analysis showed that sensor data could reliably predict subjectively reported OFF moments. These results show that our Parkinson's disease monitoring system is a feasible method to use in a diverse Parkinson's disease population for at least a period of two weeks. For longer use, the monitoring system may be too intense and wearing comfort needs to be optimized.

Introduction

Parkinson's disease (PD) is a neurodegenerative disorder characterized by motor symptoms like tremor, rigidity and bradykinesia, which can fluctuate during the day. Nowadays, symptoms and disease course are charted by a clinician and are measured most often using the MDS-UPDRS for research purposes [1]. This is however associated with limitations: (1) symptoms and disease course are assessed with low frequency and therefore only permit a snapshot of the clinical situation and may include recall bias; (2) scores given by the clinician have a subjective character; (3) patients might put themselves in a better light during in-clinical assessments compared to when they are at home. In order to get more frequent and more objective ratings of symptoms and disease course, continuous monitoring systems are essential. Such systems are believed to represent the clinical symptoms in the daily life of patients more reliably and could eventually even be used as input signal for adaptive, responsive, or closed-loop deep brain stimulation (DBS) [2].

Wearable sensors are increasingly used to detect PD symptoms due to technological innovations, resulting in small, low cost, power efficient, and accurate sensors [3–5]. Wearable sensors have shown promise in detecting tremor [6–8], freezing of gait [9], bradykinesia [10], and dyskinesia [10,11]. Recently, various monitoring systems have been developed and tested and have shown promising results [12–14]. These studies made use of predefined motor tasks in a lab or simulated home setting, which only gives a limited representation of the daily life environment. Furthermore, these regular clinical assessments cannot be performed continuously in a daily life environment.

Figure 1: Recording modalities. A: Wearable sensor attached to the wrist via a wristband. B: Screenshot of the question 'Ik ervaar tremor' ('I experience tremor') on a 7 point Likert scale, with a score of 1 indicating not at all and a score of 7 indicating very much.

The newly developed PD monitoring system presented in this paper combines wearable sensors measuring acceleration and rotational acceleration with an experience sampling method (ESM) application (Fig. 1). ESM is a validated, digital diary method consisting of multiple repeated measurements at semi-random moments in daily life [15,16]. It is superior compared to standard diary and cross-sectional assessments, because there is no recall bias and data are collected on multiple moments a day [17]. The designed ESM questionnaires include questions regarding mood states, contextual information and both motor and non-motor PD symptoms. ESM data may serve as a parallel ground truth to the wearable sensor data, making the impossible regular clinical assessments in daily life measurements redundant. The unique system combines objective data (wearable sensors) with subjective data (ESM) in the daily life of the patients.

Although the combination of wearable sensors with ESM or other electronic diary methods was often suggested in PD [18,19], and although this combination was used in other populations before [20–25], it has never been tested in PD so far. Consequently, this project aims to prove the feasibility of the new monitoring system in PD patients during daily life for two consecutive weeks. In addition, we investigated whether the ESM answers can be employed as a ground truth for the sensor data by performing an OFF moment prediction analysis. Eventually, we would like to further assess whether this system could also be used for closed-loop DBS programming.

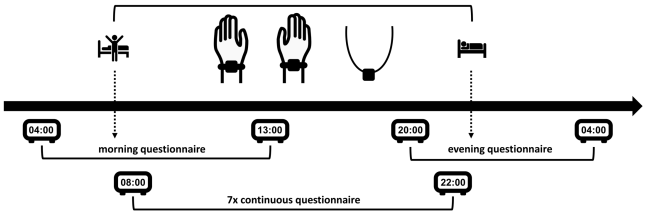


Figure 2: Schematic overview of one test day. Wearable sensors were worn from waking up until going to bed. Questionnaires were available (morning and evening) or showed up (continuous) between the indicated timeframes.

Results

Wearable sensors

On average, the group participants wore the wearable sensors 898 min a day, equalling almost 15 h. The mean time that the wearable sensors were used by the participants was 788 min (94%) within the instructed timeframe of 8.00–22.00 h (Fig. 2). Consequently, the mean non-worn time within the instructed timeframe was 52 min (6%). The mean time that the wearable sensors were used by the participants outside the instructed timeframe was 110 min.

Answers from the evaluation questionnaire showed that most participants found that the wrist (60%) and the chest (80%) sensors were comfortable to wear. However, some participants mentioned that the used accessories remained attached to clothes, or that they caused irritation. Two participants (10%) reported that the wrist sensors impaired their arm movement. According to all participants, the wrist sensors did not impair hand movement nor did the three sensors impair movements in general. None of the participants found the sensors were heavy to wear, had problems with putting the sensors on and off, or had problems with charging the sensors. Almost all participants (95%) wore the sensors in public. Some participants mentioned that they thought it was awkward to get questions about the sensors, and therefore they were glad that it was possible to wear the sensors beneath their clothes. Only 30 and 50% of the participants was willing to wear the wrist and chest sensors on a long-term basis.

Experience sampling method

ESM is a validated, digital diary method consisting of multiple repeated measurements at semi-random moments. Participants received a total of 98 continuous questionnaires, 14 morning questionnaires, and 14 evening questionnaires (Fig. 2). Only fully completed questionnaires were considered as completed. On average, 77.5 continuous questionnaires, 13.6 morning questionnaires, and 13.2 evening questionnaires were completed. This resulted in completion rates of 79.1% for the continuous questionnaires, 96.8% for the morning questionnaires and 93.9% for the evening questionnaires. For each completed continuous questionnaire, it was checked whether there was sensor data available from the two wrist sensors and the chest sensor for at least 15 min before the questionnaire was opened by the participant. We hypothesize that this timeframe will best reflect the patients clinical state belonging to the corresponding ESM

answer, and therefore this timeframe will be used for further analyses. In total, a mean of 69.0 (89.0%) continuous questionnaires were completed which had corresponding data from all three sensors (Fig. 3). For the majority of participants, only a few continuous questionnaires did not have corresponding sensor data from all three sensors for at least 15 min preceding the questionnaire (Fig. 3). Three participants (7, 10, 14 in Fig. 3) missed sensor data for more than 25% of the completed questionnaires. This was because sensors stopped recording due to full data storage as a result of hidden folders containing the removed data of previous participants. Four other participants had some aborted recordings for still unknown reasons.

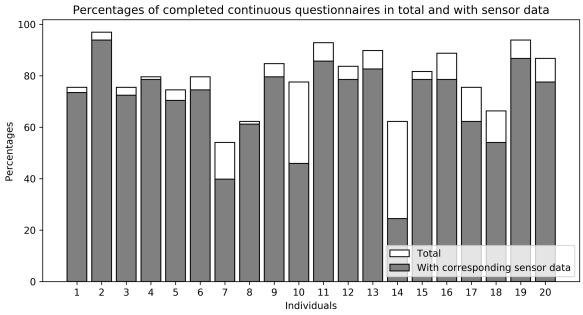


Figure 3: Completion rates. Percentages of total completed continuous questionnaires per participant (white bars) and completed continuous questionnaires with corresponding sensor data (grey bars).

Answers from the evaluation questionnaire showed that all participants found the ESM application easy to use and answered the questions independently. The majority of the participants found the ESM questions clear (85%). Some participants mentioned that the line of questioning was inconsistent. For example, a score of 1 was sometimes the most positive option and sometimes the most negative option. Three participants (15%) found it unpleasant to carry their phone with them all day, for example because they could not due to work. Further, 17 participants (85%) did not mind that the continuous questionnaires showed up at random moments during the day, which was supported by the average score of only 1.7 on the question 'I found this beep disturbing'. This item was scored on a 7 point Likert scale, where 1 corresponded with 'not disturbing at all' and 7 corresponded with 'very disturbing'. Three participants (15%) thought they missed a lot of questionnaires and ten participants (50%) said they would use the ESM application for a longer period than two consecutive weeks.

The monitoring system in general

All participants reported that the information about the research was clear, and 95% found the aim of the research clear as well. Three participants (15%) considered the study to be incriminating, and two participants (10%) adapted their daily life because of the study. For example, one participant made shorter cycling trips, since this participant was afraid to miss questionnaires.

Combining ESM and wearable sensor data

Our pilot investigation on one patient with severe ON/OFF fluctuations yielded a reliable detection of subjectively registered OFF moments based on sensor data using a logistic regression classifier (area under the curve = 0.73). Figure 4 shows true positive and false positive rates for different thresholds (Receiver- Operator-Characteristic). These initial results highlight the feasibility of using wearable sensors to detect symptom severity.

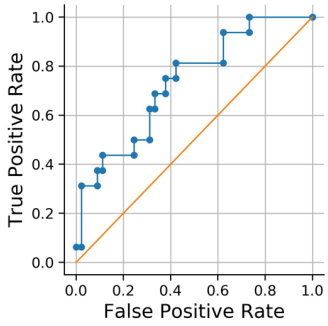


Figure 4: Receiver-Operator-Characteristic curve of detection of OFF moments from sensor data. Aurea under de curve = 0.73.

Discussion

This study showed that combining wearable sensors with ESM is a feasible method for monitoring PD patients in daily life. Results showed that participants wore the wearable sensors almost during the whole instructed timeframe and even beyond, showing that the wearable sensors were not obstructive. Furthermore, ESM completion rates were high and participants evaluated the monitoring system positively. The presented participant characteristics (Table 1) suggest that this monitoring system is feasible for a diverse population of PD patients. For example, age was ranging between 46 and 74 years, disease duration was ranging between 1 and 21 years, and no restrictions were made based on PD phenotype and treatment. There was a relatively high proportion of patients with DBS implants. This selection was most probably due to the population of the academic hospital. We cannot rule out the possibility that PD patients with DBS are more willingly to use technological devices such as this monitoring system.

ESM completion rates were high, with an average of 79%. All participants met the requirement defined in previous work, which demands that at least one third of the continuous questionnaires should be completed to have valid ESM data [26]. The high completion rate may be due to the extensive briefing during the start session and due to the phone contact moments with the participants on day 2 and 8. Compared to previous work using the same ESM application for 5 days in 5 PD patients, our completion rates were 5% lower [27]. One might argue that the longer use of the ESM application in this study might explain lower completion rates. This argument is supported by a previous N = 1 study, in which the completion rate was 47% in week 1 and dropped to 29% in week 2 [18]. We did however not see any differences between average completion rates per day (Supplementary Fig. 1). Also, there were no differences in completion rates between timeframes (Supplementary Fig. 2). The high completion rates in general are illustrated by the fact that all patients found the application easy to use, and that even half of the patients is willing to use the ESM application for a longer period.

Two participants reported they adapted their daily life during the study period. This was because participants were for example afraid to miss questionnaires. We believe that by using ESM, participants might indeed be more 'on guard' and they are required to carry their mobile phone all day. For non-digital questionnaires, there is to our knowledge no study describing whether participants adapted their daily life during the study period. We hypothesize that this will also be the case, since nondigital questionnaires mostly have to be completed at set times. Regarding completion rates it is hard to compare ESM with nondigital diary methods since non-digital diaries can be completed at different moments than instructed.

Male/female, *n*	16/4
Age, years	63.3 (7.4)
Montreal Cognitive Assessment score	27.6 (1.5)
Disease duration, years	8.1 (5.8)
Levodopa equivalent daily dose	770.4 (393.9)
Tremor/akinetic rigid, *n*	9/11
ON/OFF fluctuations yes/no, *n*	12/8
Deep brain stimulation yes/no, *n*	6/14
Deep brain stimulation duration, years	3.3 (1.5)

Table 1: **Participant characteristics.** Data are presented as means and (standard deviations).

The evaluation questionnaire outcomes showed that there is room for improvement regarding the used materials of the wearable sensor accessories. The accessories used in this study were developed with the aim that they should be easy to handle for PD patients. This was the case, since none of the participants had trouble with putting the sensors on or off. However, the use of Velcro resulted in the disadvantage that the sensors sometimes remained attached to clothes. Another disadvantage was that the sensors and accessories were not attractive enough to wear. Improvement of the accessories will likely result in a higher amount of patients who find the wearable sensors comfortable to wear and who would wear the wearable sensors on a chronic basis.

The next steps of this study are to validate ESM data, improve PD symptom algorithms for wearable data, correlate wearable data with ESM data, and eventually predict ESM scores based on wearable data. Since the combination of wearable data with ESM data is new in itself, we propose the following data processing steps: for each completed continuous ESM questionnaire, 15 minutes of sensor data prior to this completed questionnaire will be extracted. We hypothesize that this timeframe will best reflect the patients clinical state belonging to the corresponding ESM answer. It should however be tested which timeframe best fits with the questionnaire timestamp. This might be much shorter or longer than the proposed 15 min, or might even be after completion of the questionnaire. Also, it might be necessary to start with an active/inactive classification. On average it took the participants 3 min and 39 s to complete the continuous questionnaire. The start and end time of the completed questionnaires are recorded, so to ensure that the time the questionnaire was completed will not be included in the analysis. The selected timeframe will then be divided into windows of length w from which different features in the time and spectral domain will be extracted. Similar to different timeframe lengths, different window lengths for feature extraction will need to be compared as well. The extracted features can then be used for correlation of wearable data with ESM data and eventually for prediction of ESM data. See Fig. 5 for an overview of the proposed data processing pipeline.

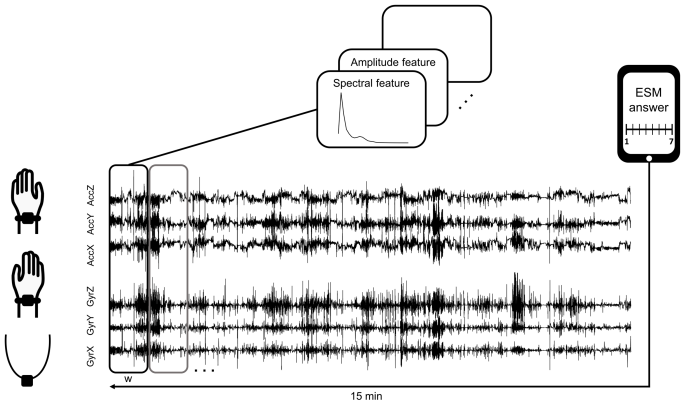


Figure 5: Proposed data processing steps. Fifteen minutes of sensor data prior to a completed questionnaire will be extracted. This timeframe will then be divided into windows of length w from which different features in the time and spectral domain will be extracted.

One limitation of the used monitoring system is that due to minimizing the number of sensors not all PD symptoms can be measured. For example, tremor in the lower limbs will likely not be measured with this system. Future studies might consider the use of smart insoles as non-obstructive sensor to measure tremor and other symptoms related to the lower limbs.

This study combined wearable sensors and ESM in PD patients. We demonstrated that our newly developed PD monitoring system is feasible and that it can be used for the continuous measurement of PD symptoms during daily life for both monitoring and treatment purposes. Further, it breaks new ground since the system collects objective wearable sensor data as well as subjective ESM data which might be used as a parallel ground truth. Therefore, this monitoring system does not require additional clinical assessments. If correlating wearable and ESM data, and eventually predicting ESM scores based on wearable data succeeds, the monitoring system can be used to monitor the patient during daily life in periods of medication changes, or periods preceding an outpatient clinic visit.

Methods

Participants

This study was approved by the medical ethical committee azM/UM and written informed consent was obtained from all participants. Twenty idiopathic PD patients participated in this study. Recruitment was done through their neurologist or neurosurgeon at the Maastricht University Medical Centre. Disease severity had to be rated as mild to severe (Hoehn and Yahr 1-4). Patients were included if they had an age between 18 and 80 years, if they were in possession of a smartphone (minimal iOS 8 or Android 4), if they could fluently speak and read Dutch, if they were available for two consecutive weeks of representative daily activities (meaning no holidays or planned hospital admission). After written informed consent, participants were tested for cognitive deficits and were excluded if they scored less than 24 points on the Montreal Cognitive Assessment. Participant characteristics are shown in Table 1. Hoehn and Yahr scores are shown in Fig. 6.

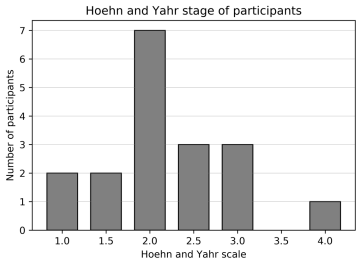


Figure 6: Disease severity of the participants indicated by Hoehn & Yahr scores.

Wearable sensors

We chose to use a new wearable, the MOX5, which was developed by the Instrument Development Engineering & Evaluation department of the Maastricht University. This sensor received CE mark approval and is available for third parties via Maastricht Instruments (Maastricht, The Netherlands). We decided to develop a wearable instead of re-using existing ones because this study acquires access to the raw accelerometer and gyroscope data, and because the on-device data storage removed the need for battery-draining data transfers. Further, this sensor has the possibility to implement symptom algorithms and online data streaming in the future.

The developed sensor contained a 6 DOF sensor, consisting of an accelerometer and gyroscope. The accelerometer covered an amplitude range of ±8 g and the gyroscope covered a range of ±2000 degrees/s. Data were collected with a sampling rate of 200 Hz. During the measurement period, each participant had to wear three wearable sensors; one at each wrist and one at the chest. The wearable sensors were attached to the body via handmade accessories (Fig. 1a). The participants had to wear the wearable sensors during daytime (preferably between 08.00 and 22.00), and had to charge them at night. Wearable sensors were aligned to the ESM questionnaires using time stamps. The participants did not have to interact with the wearable, since the measurement started as soon as the charger was removed and data was later extracted by the research team.

Experience sampling method

An ESM app, the Psymate, was downloaded and installed on the smartphone of the participants. We developed a specific ESM questionnaire, based on literature discussing relevant symptoms and items for monitoring PD at home [28,29] and based on patient and clinician interviews about relevant symptoms and items for PD monitoring at home. During the measurement period, participants were asked to complete a morning (five questions) and evening questionnaire (eight questions) which were identical on all days and which were available in the morning and evening (Supplementary Fig. 3). The morning and evening questionnaire were only available during specific timeframes and the participants were asked to complete the questionnaires when they woke up and when they were going to bed. In addition, they received a continuous questionnaire (26 questions) at seven semi randomized moments during the day (Supplementary Fig. 3). During each 2-hour block between 8.00 and 22.00 h, one continuous questionnaire was sent. See Fig. 2 for an overview of one test day. The continuous questionnaires had to be opened within 15minutes after the notification alarm and the participants were asked to complete as much questionnaires as possible without adapting their normal

daily behaviour. Participants had to rate statement questions on a Likert scale ([1–7], Fig. 1b). Some questions were multiple-choice (Supplementary Fig. 3).

Combining ESM and wearable sensor data
To investigate whether the ESM answers can be employed as a ground truth for the sensor data, we developed a prediction framework. In this framework, we used features calculated from the sensor data to predict OFF moments experienced by the patient. Based on previous literature [30–32], the following features were extracted:

1. Logarithmic Signal Energy between 3.5 and 7.5 Hz
2. Root Mean Square of the low-pass filtered (3 Hz) time series
3. Dominant Frequency and dominant energy ratio
4. Amplitude Range of the Raw Time series
5. Maximum Normalized Cross-correlation and corresponding temporal offset between all accelerometer and gyroscope channels.

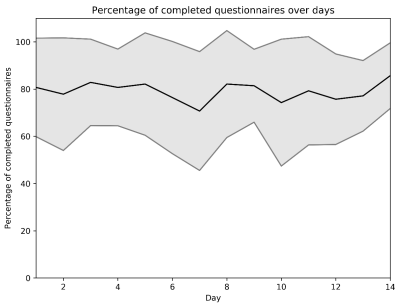
We evaluated our prediction framework in a 10-fold cross-validation and employed a simple logistic regression classifier to output probabilities for OFF moments. For this pilot study, we only evaluated one patient reporting frequent ON/OFF transitions.

Outcomes
The feasibility of our PD monitoring system will be expressed in several outcomes. Participants completed an evaluation questionnaire including questions about the use of the wearable sensors and the ESM application and about the study in general. Results of this questionnaire are outcomes for the wearable sensors, ESM and the monitoring system in general.

For the wearable sensors, the outcomes are also the minutes of collected data. This will be divided into three categories: (1) Worn time within instructed timeframe; (2) Non-worn time within instructed timeframe; (3) Worn time outside the instructed timeframe. Since the wearable sensors recorded from the moment the charger was removed until the moment the charger was plugged in again, there will be time measured in which the participant did not wear the wearable. In addition, patients may have taken off the wearable sensors during daytime when they were for example taking a shower. In order to only select the data in which the wearable sensors were actually worn, the standard deviation of the acceleration data was calculated per minute block. This was calculated for data in the x direction, which was parallel to the lower arm for the wrist sensors and which was parallel to the whole body for the chest sensor. A threshold value was empirically determined based on the standard deviation of non-worn wearable recordings. As a result, when the standard deviation per minute block was <0.002 g, and when the standard deviation of the preceding and following minute were <0.002 g as well, the minute was not included in the minutes of collected data. Minutes of collected data were averaged over the three sensors, over days, and over participants.

For the ESM, the outcomes are the percentages of completed questionnaires and the percentages of completed questionnaires for which sensor data from all three sensors was available as well. To evaluate the burden of each continuous questionnaire, we included a question on ‘how disturbing’ the corresponding questionnaire was. For the combination of ESM and sensor data, we evaluated the results of the cross-validation using the area under the curve of the Receiver-Operator-Characteristic.

Supplementary Material



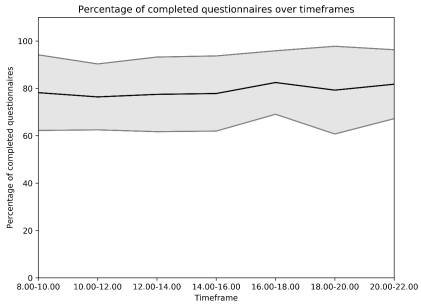


Figure S1 (upper left). Percentage of completed questionnaires over days. Light grey area is showing the standard deviation. A repeated measures ANOVA determined that completion rates did not differ significantly between days ($F(13,247) = 0.97, p = 0.49$).

Figure S2 (upper right). Percentage of completed questionnaires over timeframes. Light grey area is showing the standard deviation. A repeated measures ANOVA determined that completion rates did not differ significantly between timeframes ($F(6,114) = 1.20, p = 0.31$).

Figure S3 (right). Experience Sampling Method questionnaires.

Continuous questionnaire:

- I feel well
- I feel down
- I feel fearful
- I feel stressed
- I feel sleepy
- I am tired
- I am cheerful
- I am relaxed
- I can concentrate well
- I experience hallucinations
- I am at (home, work, travelling, at family/friends place, in public)
- I am (alone, with family, with my partner, with colleagues, with friends)
- I am doing (work, resting, household/odd jobs, sports, something else)
- I can do this without hinder
- I am comfortable walking/standing
- I can sit or stand still easily
- I can speak well
- I can walk well
- I experience tremor
- I am moving slow
- I experience stiffness
- I experience muscles tension
- I am moving uncontrollably
- I feel currently (OFF, OFF → ON, ON, ON → OFF)
- I took Parkinson medication since the last beep (yes, no, I don't recall)
- I found this beep disturbing

Morning questionnaire

- I slept well
- I woke up often last night
- I feel rested
- It was physically difficult to get up
- It was mentally difficult to get up

Evening questionnaire

- I had long OFF periods today
- I had many OFF periods today
- Walking went well today
- (un)dressing went well today
- Eating and drinking went well today
- Personal care went well today
- Household activities went well today
- I was tired today

References

1. Goetz, C. G. et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 23, 2129–2170 (2008).
2. Habets, J. G. V. et al. An update on adaptive deep brain stimulation in Parkinson's disease. Mov Disord 33, 1834–1843 (2018).
3. Rovini, E., Maremmani, C. & Cavallo, F. How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review. Front Neurosci 11, 555 (2017).
4. Sánchez-Ferro, Á. et al. New methods for the assessment of Parkinson's disease (2005 to 2015): A systematic review. Mov Disord 31, 1283–1292 (2016).
5. Thorp, J. E., Adamczyk, P. G., Ploeg, H.-L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. Front Neurol 9, 1036 (2018).
6. Basu, I. et al. Pathological tremor prediction using surface electromyogram and acceleration: potential use in 'ON-OFF' demand driven deep brain stimulator design. J Neural Eng 10, 036019 (2013).
7. Delrobaei, M. et al. Towards remote monitoring of Parkinson's disease tremor using wearable motion capture systems. J Neurol Sci 384, 38–45 (2018).
8. Khobragade, N., Graupe, D. & Tuninetti, D. Towards fully automated closed-loop Deep Brain Stimulation in Parkinson's disease patients: A LAMSTAR-based tremor predictor. Annu Int Conf IEEE Eng Med Biol Soc 2015, 2616–2619 (2015).
9. Rodríguez-Martín, D. et al. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. PLoS One 12, e0171764 (2017).
10. Griffiths, R. I. et al. Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. J Parkinsons Dis 2, 47–55 (2012).
11. Delrobaei, M., Baktash, N., Gilmore, G., McIsaac, K. & Jog, M. Using Wearable Technology to Generate Objective Parkinson's Disease Dyskinesia Severity Score: Possibilities for Home Monitoring. IEEE Trans Neural Syst Rehabil Eng 25, 1853–1863 (2017).
12. Angeles, P., Tai, Y., Pavese, N., Wilson, S. & Vaidyanathan, R. Automated assessment of symptom severity changes during deep brain stimulation (DBS) therapy for Parkinson's disease. IEEE Int Conf Rehabil Robot 2017, 1512–1517 (2017).
13. Bayés, À. et al. A 'HOLTER' for Parkinson's disease: Validation of the ability to detect on-off states using the REMPARK system. Gait Posture 59, 1–6 (2018).
14. Cancela, J. et al. Monitoring of motor and non-motor symptoms of Parkinson's disease through a mHealth platform. Annu Int Conf IEEE Eng Med Biol Soc 2016, 663–666 (2016).
15. Csikszentmihalyi, M. & Larson, R. Validity and reliability of the Experience-Sampling Method. J Nerv Ment Dis 175, 526–536 (1987).
16. Myin-Germeys, I. et al. Experience sampling research in psychopathology: opening the black box of daily life. Psychol Med 39, 1533–1547 (2009).
17. Palmier-Claus, J. E. et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 123, 12–20 (2011).
18. van der Velden, R. M. J., Mulders, A. E. P., Drukker, M., Kuijf, M. L. & Leentjens, A. F. G. Network analysis of symptoms in a Parkinson patient using experience sampling data: An n = 1 study. Mov Disord 33, 1938–1944 (2018).
19. Vizcarra, J. A. et al. The Parkinson's disease e-diary: Developing a clinical and research tool for the digital age. Mov Disord 34, 676–681 (2019).
20. Bedard, C. et al. Understanding Environmental and Contextual Influences of Physical Activity During First-Year University: The Feasibility of Using Ecological Momentary Assessment in the MovingU Study. JMIR Public Health Surveill 3, e32 (2017).
21. Booij, S. H. et al. Cortisol and α-Amylase Secretion Patterns between and within Depressed and Non-Depressed Individuals. PLoS One 10, e0131002 (2015).
22. Brannon, E. E., Cushing, C. C., Crick, C. J. & Mitchell, T. B. The promise of wearable sensors and ecological momentary assessment measures for dynamical systems modeling in adolescents: a feasibility and acceptability study. Transl Behav Med 6, 558–565 (2016).
23. Knell, G. et al. Ecological Momentary Assessment of Physical Activity: Validation Study. J Med Internet Res 19, e253 (2017).

24. Liao, Y., Chou, C.-P., Huh, J., Leventhal, A. & Dunton, G. Associations of Affective Responses During Free-Living Physical Activity and Future Physical Activity Levels: an Ecological Momentary Assessment Study. Int J Behav Med 24, 513–519 (2017).
25. Maher, J. P., Dzubur, E., Huh, J., Intille, S. & Dunton, G. F. Within-Day Time-Varying Associations Between Behavioral Cognitions and Physical Activity in Adults. J Sport Exerc Psychol 38, 423–434 (2016).
26. Delespaul, P. a. E. G. Assessing schizophrenia in daily life : the experience sampling method. (1995).
27. Broen, M. P. G. et al. Unraveling the Relationship between Motor Symptoms, Affective States and Contextual Factors in Parkinson's Disease: A Feasibility Study of the Experience Sampling Method. PLoS One 11, e0151195 (2016).
28. Ferreira, J. J. et al. Clinical Parameters and Tools for Home-Based Assessment of Parkinson's Disease: Results from a Delphi study. J Parkinsons Dis 5, 281–290 (2015).
29. Serrano, J. A. et al. Participatory design in Parkinson's research with focus on the symptomatic domains to be measured. J Parkinsons Dis 5, 187–196 (2015).
30. Hoff, J. I., Wagemans, E. A. & van Hilten, B. J. Ambulatory objective assessment of tremor in Parkinson's disease. Clin Neuropharmacol 24, 280–283 (2001).
31. Patel, S. et al. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. IEEE Trans Inf Technol Biomed 13, 864–873 (2009).
32. Salarian, A. et al. Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system. IEEE Trans Biomed Eng 54, 313–322 (2007).

Chapter 6

Mobile Health Daily Life Monitoring for Parkinson Disease: Development and Validation of Ecological Momentary Assessments

Habets JGV, Heijmans M, Herff C,
Simons CJP, Leentjens AFG,
Temel Y, Kuijf ML, Kubben PL

Published in JMIR Mhealth and Uhealth 2020, 8(5)
DOI: https://doi.org/10.2196/15628

Abstract

Introduction: Parkinson's disease (PD) monitoring is making a transition from periodic clinical assessments to continuous daily life monitoring in 'free-living' conditions. Traditional PD monitor methods lack intraday fluctuation detection. Electronical diaries (eDiaries) hold potential to collect subjective experiences on the severity and burden of (non-)motor symptoms in free-living conditions.
Objective: We aim to develop a PD specific eDiary based on ecological momentary assessments (EMA) and explore its validation.
Methods: An observational cohort of twenty PD patients used the smartphone-based EMA eDiary for fourteen consecutive days without adjusting free-living routines. It presented an identical questionnaire consisting questions regarding affect, context, motor and non-motor symptoms and motor performance seven times daily at semi-randomized moments. Additionally, patients were asked to complete a morning and an evening questionnaire.
Results: Mean affect correlated respectively moderate to strong and moderate with motor performance ($R = 0.38 – 0.75$, $p<0.001$) and motor symptom ($R = 0.34 – 0.50$, $p<0.001$) items. Motor performance showed a weak to moderate negative correlation with motor symptoms ($R = -0.31 - -0.48$, $p<0.001$). Group mean answers given in on- versus wearing off-medication conditions differed significantly ($p < 0.05$), however not enough questionnaires were completed in wearing off condition to reproduce these findings on individual levels.
Conclusions: We present a PD specific EMA-eDiary. Correlations between given answers support the internal validity of the eDiary and underline EMA's potential in free-living PD monitoring. Careful patient selection and EMA design adjustment to this targeted population and their fluctuations are necessary to generate robust proof of EMA validation in future work. Combining clinical PD knowledge with practical EMA experience is inevitable to design and perform studies which will lead to successful integration of eDiaries in free-living PD monitoring.

Introduction

Parkinson disease (Parkinson Disease) is a neurodegenerative disorder that is characterized by bradykinesia, rigidity, and tremor. Many patients develop fluctuations in cardinal motor symptoms, such as bradykinesia, tremor, and postural instability, and levodopa-induced dyskinesia [1,2]. Nonmotor symptoms may also show fluctuations during the day [3,4]. Current gold standards in symptom monitoring, such as the Movement Disorders Society (MDS)—Unified Parkinson's Disease Rating Scale and the Parkinson's Disease Quality of Life-39, are suboptimal to detect such fluctuations over short periods, as they cover a longer temporal domain and require active observed tasks [5,6]. Monitoring methods that can also detect motor and nonmotor fluctuations over shorter periods in free-living conditions can contribute to applying personalized medicine in Parkinson Disease [7,8]. Examples of such new methods are telemonitoring [9] and mobile health (mHealth) apps, often including wearable sensor [10-12]. Patients with neurological conditions are believed to be able to use mobile apps [13]; however, the quality, validation, and usability of the available apps are often low [14]. Nonetheless, there have been promising results of using mHealth monitoring systems for Parkinson Disease motor and nonmotor symptoms during free-living situations [15-17].

Electronic diaries (eDiaries) hold the potential to contribute to these new monitoring methods by collecting valuable information on motor symptoms [18,19] and non-motor symptoms in free-living conditions [16].

Recently published recommendations on Parkinson Disease electronic diary (eDiary) development by a specific MDS Task Force and Committee underline the relevance and potential of this approach [20,21]. Ecological momentary assessment (EMA), also referred to as an experience sampling method, is a method that collects subjective experiences at multiple, semirandomized moments during a day. Commonly used in psychiatric and psychological populations, it holds the potential for somatic diseases as well [22]. The scarce literature describing EMA in Parkinson Disease reports feasibility in small cohorts of up to 5 patients. Reproduction and further investigation of the usefulness and value of EMA in Parkinson Disease are needed [4,16,23].

We developed the first specific Parkinson Disease eDiary using EMA and set the first steps to validate the EMA method in a broad Parkinson Disease cohort.

Methods

Study Population

We included 20 patients who were diagnosed with Parkinson Disease following the UK Parkinson's Disease Society Brain Bank Diagnostic Criteria, who were aged between 18 and 80 years, who possessed a smartphone (at least Android 4 or iPhone operating system 8), and who had adequate proficiency in the Dutch language. A Montreal Cognitive Assessment scale score lower than 24 was the only exclusion criterion [24]. Demographic and general disease characteristics were collected, such as sex and age, Parkinson Disease duration, levodopa equivalent daily dosage

(LEDD), number of daily dopaminergic medication intake moments, presence of intraday motor fluctuations, and recent Hoehn and Yahr scores.

Ecological Momentary Assessment Study Design

Participants enrolled between August 2018 and March 2019 and participated for 14 consecutive days. EMA questionnaires (referred to as beeps) were presented at seven semirandomized moments a day, one beep within every block of 2 hours between 8 AM and 10 PM. The questionnaire had to be opened within 15 min after notification to prevent procrastination. A separate morning questionnaire was available between 4 AM and 1 PM, and an evening questionnaire was available between 8 PM and 4 AM. Answers on statement questions were given on a 7-point Likert scale. The EMA method was executed via the smartphone app, PsyMate [25]. EMA was combined with the use of three wearable sensors containing accelerometers and gyroscopes. Technical details of the protocol design and feasibility analyses are reported earlier [26]. The study protocol was conducted following the Helsinki guidelines and was approved by the local medical ethical committee of Maastricht UMC+.

Data Preparation

Patients with a completion rate lower than 33% were excluded from analyses [27]. Beeps containing missing values because of unfinished questionnaires or digital data transmission failure were excluded.

To analyze positive and negative affect, we calculated the mean of the items *feeling well*, *feeling cheerful*, and *feeling relaxed* and the mean of the items *feeling down*, *feeling fearful*, and *feeling stressed*, respectively. To analyze general motor function, we calculated the mean of the items *ability to perform current activity*, *ability to walk well*, *ability to talk well*, and *to experience steady mobility*. When we refer to the items *mean positive affect*, *mean negative affect*, or *general motor function* in the paper, we are referring to these calculated mean scores. General motor function in the evening questionnaire was calculated as the mean of the evening questionnaire items *ability to dress*, *ability to eat*, *ability to do household activities*, *ability to do personal care*, and *ability to walk*. The evening questionnaire items *experienced many off periods* and *experienced long off periods* were averaged in an item representing off-moment severity during the day.

To represent the change in an item since the last beep, we calculated differences over time scores. The answer to the previous beep (t−1) was subtracted from the answer of the current beep (t). Two beeps are, on average, 2 hours separated from each other. We did not calculate the difference in scores between the first completed beep a day and the last beep of the previous day.

Statistical Analysis

We analyzed means, standard deviations, and distributions per item. A skewed distribution of answers of an item to the minimum (1) or the maximum (7) is called a floor or a ceiling effect, respectively. If present, we evaluated whether this floor or ceiling effect could be expected and

could be accepted or might be based on an invalid, nonspecific, or nonsensitive question and deserved further evaluation.

To validate whether items measure what they are intended to measure, the correlation between an item and a gold standard that measures the same concept can be assessed. If this expected correlation is present, this means the construct validity of that item is proven [20,28]. As there are no validated assessment scales that assess Parkinson Disease symptoms as frequent as our EMA beeps, there is a lack of a gold standard measure. Therefore, we assessed the construct validity by analyzing correlations between items from the same beep that are expected to correlate based on clinical knowledge. To further analyze construct validity, we analyzed correlations between the mean answer over all beeps during 1 day and the answer from the corresponding evening questionnaire. For the latter, we excluded days without the completed evening questionnaire. As the theoretically expected correlation of sleep with other symptoms is ambiguous, we excluded the morning questionnaires from validation analyses.

We compared beep answers given in different medication conditions to explore differences in symptom severity. We merged the two transition conditions, from on-medication to off-medication and vice versa, to differentiate three conditions: on-medication condition, off-medication condition, and the transition between on- and off-medication condition.

By calculating correlations between scores of items that are expected to correlate, we analyzed the sensitivity of our EMA questionnaire to measure changes over time. We explored the differences between beep answers given in different medication conditions, on-medication condition, off-medication condition, and transitions between the two. The beeps identified as off-medication condition represent the wearing-off medication condition because the patients were never fully depleted of dopaminergic medication. In the rest of the paper, we will use the term on-beeps and off-beeps to refer to these medication conditions during a completed beep questionnaire. We performed these comparisons on group and individual levels. The significance of differences between the different medication conditions was calculated using Mann-Whitney U tests. Correlations were calculated using Spearman correlation tests. *P* values were corrected with a Bonferroni correction. All the data preparation and statistical analyses were performed in Python Jupyter Notebook 3 using packages pandas (version 0.24.2), Numpy (version 1.16.4), datetime (version 1.0.0), and Scipy (version 1.3.0).

Results

Study Population

We included 4 female and 16 male patients with idiopathic Parkinson Disease with a mean age of 63 years (SD 7), a mean disease duration of 8 years (SD 6), and a mean LEDD of 770 mg (SD 394); 6 participants were treated with deep brain stimulation for a mean period of 3.3 years (SD 1.5; Table 1). The mean completion rate was 78% out of 98 continuous beeps (SD 12). No participants were excluded based on a too low completion rate (ie, completion rate <33%) [27].

Gender ratio (female:male)	**4:16**
Age (years), mean (SD)	63 (7)
Disease duration (years), mean (SD)	8 (6)
Levodopa equivalent daily dosage (mg), mean (SD)	770 (394)
DBS[a]	
Patients with DBS treatment, n	6
Duration of DBS treatment (years), mean (SD)	3.3 (1.5)
H&Y[b], n (%)	
1	2 (10)
1.5	2 (10)
2	7 (35)
2.5	3 (10)
3	3 (15)
4	1 (5)
Montreal Cognitive Assessment, mean (SD)	27.6 (1.5)

Table 1: demographics of study population. All durations are in years. DBS: deep brain stimulation, H&Y: Hoeh and Yahr scale, mg: milligrams, MoCA: Montreal Cognitive Assessment, sd: standard deviation.

Parkinson specific EMA Questionnaire Development

Affect and context items from widely applied EMA questionnaires in psychiatry were added [29]. Parkinson Disease-specific items are based on a literature search and structured interviews with clinicians, patients, and caregivers. A detailed description of this literature search and the structured interviews can be found in the Supplementary Material.

Repeated discussions with the *EMA expert group* in our institution (among them CS) gave us the following insights into designing a valid EMA questionnaire for patients with Parkinson Disease: (1) do not only assess motor symptoms by direct questions about the specific motor symptom, (2) include assessment of the burden or the influence of the symptoms on the patient's performance/well-being, and (3) include items on context (where/with whom/what) and affect and to have the possibility to correct for varying settings or mood fluctuations. On the basis of the advice of the EMA expert group, we consistently phrased the questions as statements in the "I" perspective and tried to avoid confirming ("I do feel...") and denying ("I do not feel...") statements next to each other [27,30]. Furthermore, when translating clinical terms or items from retrospective questionnaires into EMA items, we aimed to maximize face validity by using everyday language. The final EMA questionnaire is shown in Figure 1.

Beep questionnaire (semi-random repeated moments)	**Morning questionnaire**
I feel well	I slept well
I feel down	I woke up often last night
I feel fearful	I feel rested
I feel stressed	It was physically difficult to get up
I feel sleepy	It was mentally difficult to get up
I am tired	
I am cheerful	
I am relaxed	
I can concentrate well	
I experience hallucinations	**Evening questionnaire**
I am at [home, work, travelling, at family/friend's place, in public]	I had long OFF periods today
I am with [nobody, family, partner, colleagues, friends]	I had many OFF periods today
I am doing [work, resting, household/odd jobs, sports, something else]	Walking went well today
I can do this without hinder	(un)dressing went well today
I am comfortable walking/standing	Eating/ drinking went well today
I can sit or stand still easily	Personal care went well today
I can speak easily	Household activities went well today
I can walk easily	I was tired today
I experience tremor	
I am moving slow	
I experience stiffness	
My muscles are tensioned	
I am uncontrollable moving	
I feel in medication state … [1: OFF, 2: ON -> OFF, 3: ON, 4: OFF -> ON]	
I took Parkinson medication since last beep [yes, no, I don't recall]	

Figure 1: PD EMA questionnaire content, English translation from original Dutch version. The beep questionnaire which is presented seven times during the day represents the four motor domains, as well as affect, cognition, context and motor performance. The evening questionnaire covers off moments and motor performance over the day and the morning questionnaire covers sleep.

Parkinson Disease Ecological Momentary Assessment Questionnaire Validity

Figure 2 shows means and distributions of all participants per item from the beep questionnaire and the evening questionnaire. Positive affect shows a small ceiling effect, whereas negative affect shows a floor effect. Experiencing hallucinations shows a strong floor effect, with only one participant experiencing hallucinations. Positive formulated items on motor functioning show a small floor effect. Experiencing tremor and dyskinesia shows stronger floor effects than experiencing slowness and stiffness.

Construct validity was assessed by evaluating the presence of expected correlations between items (Figure 3). Mean positive and negative affect showed a strong negative correlation with each other (R=−0.71; *P*<.001; Figure 3). Both positive and negative affect scores showed moderate-to-strong correlations with general motor functioning (R=0.75 and R=−0.53, respectively; *P*<.001). Mean positive and negative affect scores showed weak-to-moderate

correlations with different motor symptoms (R=−0.37 to −0.49 and R=0.32 to 0.50, respectively; *P*<.001). General motor functioning showed moderate-to-weak correlations with the motor symptoms tremor, slowness, stiffness, and dyskinesia (R=−0.34, −0.47, −0.44, and −0.51, respectively; *P*<.001).

Beep answers on mean affect scores and general motor functioning from 1 day showed moderate correlations with both the items assessing the amount of experienced off-beeps and the general motor performance from the corresponding evening questionnaires in expected directions (R=−0.43 to 0.69; *P*<.001). Beep answers during the day on slowness, stiffness, tremor, and dyskinesia showed weak-to-moderate correlations with general motor functioning answers from the evening questionnaire (R=−0.24 to 0.44; *P*<.001). These items assessing motor symptoms in the beep questionnaires, also showed weak-to-moderate correlations in the expected directions with the item assessing off-beeps in the evening questionnaire. (R=0.24 to 0.69; *P*<.001). Although dyskinesia is no typical symptom during off-beeps, it correlated strongly with off-beeps over the whole day (R=0.69; *P*<.001).

The correlations between difference over time scores were less strong as the absolute answers (see Supplementary Material for a correlation heatmap of difference over time scores). All correlations were weak to absent.

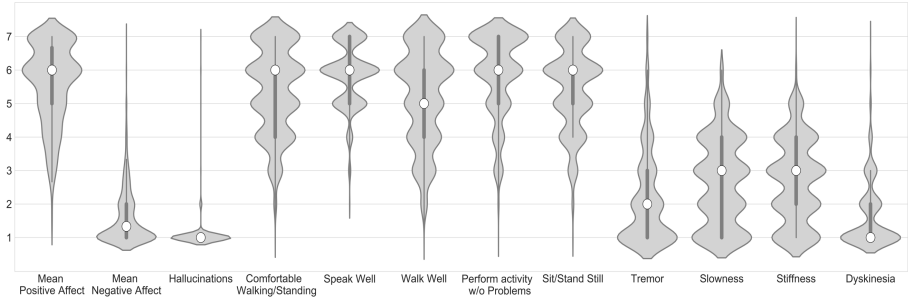


Figure 2: Distribution plots of answers from beep questionnaires. Mean positive and negative affect showed high and low mean answers, respectively. The ability to perform daily life tasks showed moderate-to-high mean answers, whereas the motor symptom items showed low-to-moderate mean answers. All items were statements and were answered on a 7-point Likert scale, ranging from 1 (not at all) to 7 (very). The white dot represents the median answer, the thick black line represents the IQR, and the thin black lines represent the rest of the distribution, calculated as IQR times 1.5. The width of the shapes correlates with the probability that the patient answered the corresponding value.

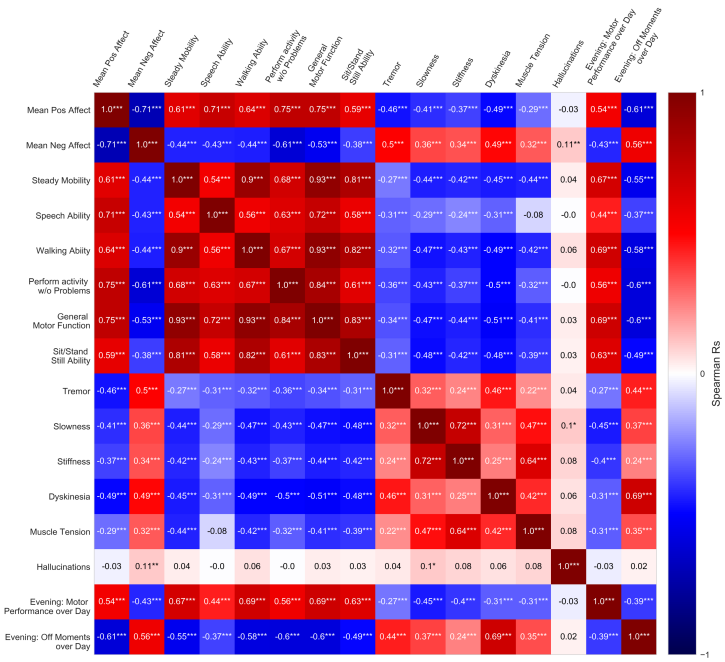


Figure 3: Correlations between items from the beep questionnaire and the evening questionnaires. We observed strong and moderate correlations between the motor performance items and mean positive and negative affect, respectively. We observed correlations between mean affect scores, motor symptoms, motor performance, and medication states, in the beep questionnaires and the evening questionnaires, in directions that were expected**.**

Spearman's correlations are shown for the corresponding items. Significance level is indicated with * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$. Bonferroni permutation correction is performed. The items from the evening questionnaire are correlated with mean answers given on the corresponding day. Absolute R correlation values lower than 0.15 are colored black for readability.

Influence of Medication Condition on EMA Answers

Of 1573 beeps, 1195 (75.97%) were labeled by patients as *answered in on-medication condition* (on-beeps), 339 beeps (21.56%) were labeled by patients as *answered in between on- and off-medication condition* (transition beeps), and 39 beeps (2.48%) were labeled by patients as *answered in off-medication condition* (off-beeps; Figure 4). On a group level, mean answers significantly differed between on-beeps and non–on-beeps for mean positive affect, general motor function, slowness, and dyskinesia. Mean answers between on-beeps and off-beeps significantly differed for mean positive affect, mean negative affect, general motor function, and tremor.

On an individual level, mean answers during different medication conditions did not significantly differ. The differences were either not significant or not relevant. Only 5 of 20 participants reported 20% or more beeps in the *non–on-medication condition* (Supplementary Material).

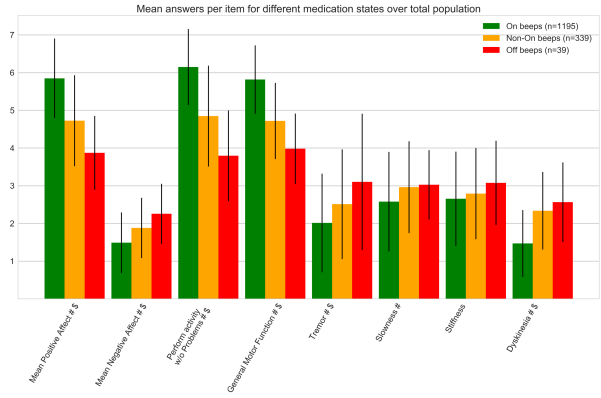


Figure 4: Mean answers during different medication states. The given answers in different medication conditions show significant differences. The direction of differences is as expected, except for dyskinesia, which is scored higher, on average, during off-condition compared with on-condition. Whiskers indicate standard deviations. On-beeps represent answers during on-medication, non–on-beeps represent answers during off-medication and during the transition phase, and off-beeps represent answers only during off-medication. #: significant difference ($p < 0.05$) between on-beeps and non-on-beeps. $: significant difference ($p < 0.05$) between on-beeps and off-beeps.

Discussion

Clinical Relevance of EMA for Free-Living Parkinson Monitoring

Owing to the fluctuating nature of Parkinson Disease and its heterogeneous character, EMA holds theoretically great potential to increase insight into symptom severity and burden fluctuation during free-living conditions. Our work fits in the first milestone defined by the MDS Technology Task Force and the MDS Rating Scales Program Electronic Development Ad-Hoc Committee by giving insight into the prioritization of outcomes, which are relevant for the patient to measure [20]. Obviously, this paper is one of the first where many are to follow.

A common challenge for all future work in this field is the validation of methods and questionnaires. Validated scales only exist for Parkinson Disease monitoring with longer time intervals than the short time intervals needed to detect intraday fluctuations. This fact makes classical validation with golden standards hard and even incorrect depending on the methodology. The MDS Task Force on Technology, therefore, advices to validate new Parkinson Disease monitoring methods for free-living conditions according to accuracy, reliability, sensitivity, and minimal clinically significant differences [21]. This validation challenge is also relevant for the integration of additional biometric monitor devices, such as accelerometers, gyroscopes, microphones, or electrophysiological monitor devices. Vizcarra et al [20] make a distinction between the integration of action-dependent and action-independent monitoring. It is expected that creating golden standards for action-dependent tasks in, for example, a laboratory setting is easier than creating standards for an action-independent setting such as free-living [18,19]. For the latter, validated Parkinson Disease monitoring devices collecting subjective experiences on symptom severity and burden can be of substantial value.

The most applied and promising action-independent Parkinson Disease monitoring methods for free-living conditions are based on wearable sensors [31,32]. Attempts to include subjective diary data in the validation of sensor data algorithms were hindered by practical limitations mainly, for example, recall bias and diary fatigue [19,33]. Smartphone-based EMA methodology can be applied less obtrusively and address these traditional diary limitations. Naturally, the feasibility of this method is heavily dependent on the frequency and duration in which it is applied. These factors require thorough future investigations and may differ per intention of use, for example, wearable sensor calibration or periodic free-living monitoring of nonmotor symptoms.

General Lessons on EMA in Parkinson's Disease

The introduction of a new method in Parkinson Disease monitoring entails challenges and questions beyond the current literature. To address these challenges and questions as good as possible, we gathered a multidisciplinary team consisting of clinical Parkinson Disease expertise (neurology, specialized nurses, neuropsychiatry, and neurosurgery) and experienced practitioners of EMA (neuropsychology and neuropsychiatry). We described our most important lessons regarding the content and the phrasing of the EMA questionnaire to inform clinicians and researchers interested in applying EMA in Parkinson Disease. Moreover, a recently published checklist provides researchers with a tool to design an EMA-based diary study [34]. Essential for EMA in Parkinson Disease is the similarity between the frequency of EMA assessments and the frequency of symptom fluctuations that are intended to capture. Thus, EMA studies may require different designs depending on whether they monitor levodopa-induced dyskinesia fluctuations over a day or whether they monitor the effect of an extra levodopa agonist on morning bradykinesia.

Validation of EMA in Parkinson's Disease

Mean answer values and distributions show expected findings (Figure 2). Positive affect items are known to be answered higher than negative affect items [35]. The high mean answers on general motor function and the low mean answers on motor symptoms can be explained by the stable-treated population and the relatively low overall disease progression (Table 1). Concerning the observed floor and ceiling effects, we only regard the item on hallucinations as obsolete for this population because of the observed extreme floor effect. As stated earlier, negative affect items are known to show a floor effect. Tremor and dyskinesia also show an unsatisfying floor effect, although we think this is because of the low prevalence of these symptoms in our sample. Moreover, the unexpected positive correlation between dyskinesia and experienced off-beeps suggests that the dyskinesia item might not be well understood by patients. Limited awareness on the presence of dyskinesia among patients with Parkinson Disease is described earlier [36]. This finding might also be strengthened by the population's low prevalence of dyskinesia. We advise, therefore, to avoid the use of nonapplicable, general questions for individual patients. If an item is not applicable for a patient, the patient should be clearly instructed on how to answer this item.

The moderate-to-high correlations present between affect, motor function, and motor symptoms prove the construct validity of the Parkinson Disease EMA method partially. The low-to-moderate correlations between motor function and motor symptoms warrant cautious conclusions, and follow-up validation among a narrower selected population with more motor fluctuations is needed to more extensively proof construct validity.

The high number of beeps answered in on-medication condition (Figure 4) and the weak till absent correlations between difference over time scores (see Supplementary Material) confirm this hypothesis. Significance levels are calculated using Mann-Whitney U tests (all $P<0.5$). All questions except *Stiffness* differed significantly between on-beeps and non–on-beeps. All questions except *Slowness* and *Stiffness* differed significantly between on-beeps and off-beeps. Ideally, the significant differences that were only found on group level also hold on individual levels in the next validation study, especially because EMA is intended for individual monitoring.

Despite the fact that further investigation is needed, EMA in Parkinson disease seems to be potentially useful and valid when evaluating the moderate-to-high correlations between affect, general functioning, bradykinesia, and stiffness. Altogether, we interpret our findings as encouraging, and we stress the importance of a careful patient selection depending on the exact goal of EMA monitoring.

Limitations

The broad inclusion policy was a well-considered choice in the study design, and it resulted in important information about the feasibility and validity of EMA in a broad Parkinson Disease population. When applied in a more specified cohort, clinimetric validation analyses necessary for the next step in validation are better feasible, such as principal component analyses to exclude fewer sensitive items. The latter may lead to individual patient- or patient subgroup–specific questionnaire content.

Conclusions

EMA-based eDiaries are promising to enrich free-living Parkinson Disease monitoring with essential information on motor and nonmotor fluctuations. First validation analyses suggest the internal validation of EMA among a general Parkinson Disease population. Careful patient selection and EMA design adjustment to this targeted population and their fluctuations are necessary to generate robust proof of EMA validation in future work. Combining clinical Parkinson Disease knowledge with practical EMA experience is inevitable to design and perform studies, which will lead to successful integration of eDiaries in free-living Parkinson Disease monitoring.

Acknowledgements

The authors would like to thank Mirella Waber for advice on the application and content of the eDiary among patients with Parkinson Disease, the EMA expert group within the Department of Psychiatry and Neuropsychology at Maastricht University for constructive discussion on the EMA methodology, and Karel Borkelmans for the technical support of the PsyMate app. YT and PK received a grant from Stichting Weijerhorst.

None conflict of interest to report.

Supplementary Material

Development of Parkinson's disease (PD) specific Ecological Momentary Assessment (EMA) list
To determine the content of a PD EMA questionnaire we performed a literature search, structured interviews with clinical experts, patients and caregivers and consulted the EMA expert group within our institution.
We searched the Pubmed database in February 2018 with the following search strategy: '*(("Ecological Momentary Assessment"[Mesh]) OR ((experience sampling method) OR (ESM) OR (EMA) OR (ecological momentary assessment))) AND (("Parkinson Disease"[Mesh]) OR ("Movement Disorders"[Mesh]) OR parkinson* OR (movement disorder*))*'. This resulted in 47 hits, of which one described EMA in PD.[16] Even this study did not use a specialized EMA questionnaire for PD. Next, we reviewed literature describing relevant aspects for PD monitoring at home. The SENSE-PARK research group published two well-conducted studies on relevant parameters in PD monitoring at home. To define what characterizes 'good' and 'bad' Parkinson moments at home, they performed a web-survey among 198 patients followed by 6 focus groups, and performed a Delphi-study among 12 clinicians. They extracted six domains to monitor: gait, bradykinesia, tremor, sleep, sway, cognition.[37] To define which parameters and assessment tools are needed to monitor these six domains, they performed a 2-round Delphi-study among 12 clinicians, 159 patients and 72 caregivers.[38]
Ferreira et al asked PD patients, caregivers and clinicians to score the importance of parameters in PD monitoring in the home situation between 1 and 5, and they asked them to rate a top 3 most important parameters. Here, we present a list of parameters that are scored higher than 4.0 on average and are mentioned in the top 3 parameters in more than 33% of the subjects. They are presented in the second column of table S1.

We performed twenty structured interviews with individual patients and asked them '*What complaint distinguishes a good from a bad Parkinson moment the most?*' and '*What restriction that limits you in daily life distinguishes a good from a bad Parkinson moment the most?*'. We performed six structured interviews with movement disorders experts within our academic hospital, and asked their opinion about the same questions. Results on how many percent of patients and clinical experts mentioned symptoms to monitor are shown in table S1.

	SENSE-PARK research group		**Own structured interviews**	
Parameter	Serrano et al[37]*	Ferreira et al[38]**	% of patients	% of experts
stiffness/ slowness	1.7/ 2.2		80	83
walking		4.1, 43%	45	100
Freezing of gait		4.3, 59%		
dyskinesia	0.1		40	67
tremor	1.5	4.1, 36%	40	100
tired	1.5		35	33
executive functioning	0.6	4.7, 98%	25	33
clothing/washing		4.7, 91%	25	83
housekeeping		4.7, 91%	25	50

balance/falls	1.0	4.6, 81%	20	83
fine motor movements	1.7	4.5, 55%	20	33
speech	0.5		15	83
pain (legs)	1.0		15	17
activities		4.7, 91%	10	83
writing		3.9, 43%	5	0
autonomous	1.9		5	17
ON vs OFF			0	50
Sleep		4.4, 66%	0	0

Table S1: Overview of results of different authors on domains and parameters of importance in PD monitoring in the home environment

*: Serrano et al present their results as a combined score between the open survey round and the focus group results. **: Ferreira et al present their results as the average score (between 1 and 5) and the percentage the parameter was rated in the top 3 parameters to monitor.

As described in the manuscript, we distilled specific questions about (motor) performance and (motor) symptoms based on the results shown in table S1.

Sensitivity to change analyses with difference over time scores

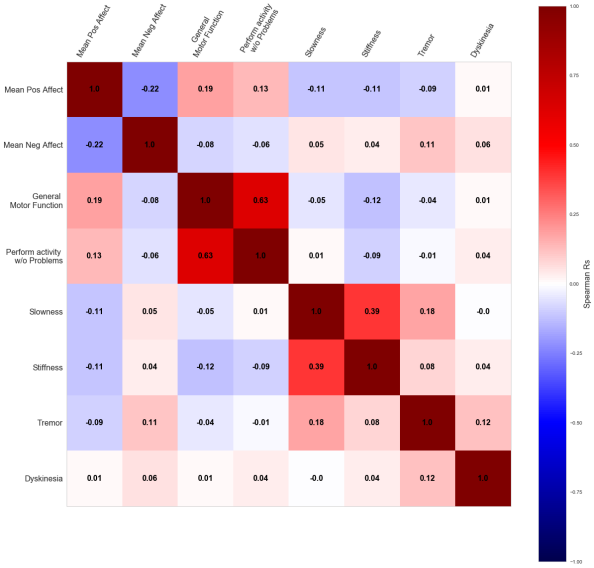


Figure S1: Correlations between difference over time scores

Changes in mean positive affect were weakly correlated with changes in mean negative affect (R = -0.22). Mean positive and negative affect changed parallel with the motor symptoms in the expected directions, although the correlations were very weak till weak (R = 0.01 – 0.22). Changes since the previous beep in general function only correlated weakly with mean positive affect (R = 0.19) and were not correlated with the rest of the items. Changes in slowness and stiffness did correlate with each other (R = 0.39) and correlated weakly with changes in tremor (R = 0.18 and 0.08).

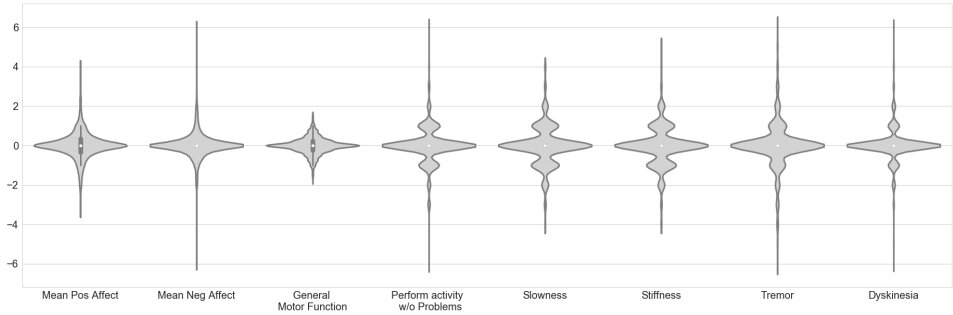


Figure S2: Distribution plot for difference over time scores of most relevant items. The very high relative amount of (nearly) 0 answers means that the answers on these items did not show large differences compared to the answers on the same items of the prior completed questionnaire. The percentage of beeps with a difference over time score of '0' was 45% for mean positive affect, 55% for mean negative affect, 41% for general motor functioning, 57% for slowness, 56% for stiffness, 65% for tremor and 76% for dyskinesia.

Detailed analysis of answers given during different medication conditions

ID	reported_fluct	ON beeps	non-ON beeps	OFF beeps	mean(sd)_OnOffnew	mean(sd)_many_offs	mean(sd)_long_offs
110001	No	74 (96.1%)	3 (3.9%)	0 (0.0%)	3.0 (0.2)	1.0 (0.0)	1.0 (0.0)
110002	Yes	20 (20.8%)	76 (79.2%)	20 (20.8%)	2.0 (0.6)	4.0 (0.8)	3.9 (0.8)
110003	Yes	72 (97.3%)	2 (2.7%)	0 (0.0%)	3.0 (0.2)	1.1 (0.3)	1.0 (0.0)
110004	Yes	35 (44.9%)	43 (55.1%)	2 (2.6%)	2.4 (0.5)	3.3 (1.5)	2.4 (1.5)
110005	No	73 (100.0%)	0 (0.0%)	0 (0.0%)	3.0 (0.0)	1.2 (0.4)	1.2 (0.4)
110006	Yes	62 (79.5%)	16 (20.5%)	0 (0.0%)	2.8 (0.4)	2.4 (1.7)	2.4 (1.5)
110007	No	51 (96.2%)	2 (3.8%)	0 (0.0%)	3.0 (0.2)	3.8 (1.1)	3.2 (1.3)
110008	Yes	6 (9.8%)	55 (90.2%)	1 (1.6%)	2.1 (0.3)	1.5 (1.1)	1.5 (1.0)
110009	No	58 (93.5%)	4 (6.5%)	2 (3.2%)	2.9 (0.4)	1.6 (1.3)	1.1 (0.3)
110010	No	75 (97.4%)	2 (2.6%)	0 (0.0%)	3.0 (0.2)	1.0 (0.0)	1.0 (0.0)
110011	No	91 (100.0%)	0 (0.0%)	0 (0.0%)	3.0 (0.0)	1.1 (0.3)	1.0 (0.0)
110013	No	82 (100.0%)	0 (0.0%)	0 (0.0%)	3.0 (0.0)	1.0 (0.0)	1.0 (0.0)
110014	Yes	82 (93.2%)	6 (6.8%)	0 (0.0%)	2.9 (0.3)	2.1 (1.5)	2.4 (1.7)
110015	No	59 (96.7%)	2 (3.3%)	0 (0.0%)	3.0 (0.2)	1.4 (0.5)	1.2 (0.4)
110016	Yes	79 (98.8%)	1 (1.2%)	0 (0.0%)	3.0 (0.1)	3.8 (1.3)	3.1 (1.4)
110017	Yes	19 (21.8%)	68 (78.2%)	5 (5.7%)	2.2 (0.5)	4.0 (0.0)	3.9 (0.4)
110018	Yes	25 (33.8%)	49 (66.2%)	8 (10.8%)	2.2 (0.6)	3.5 (0.6)	3.3 (1.1)
110019	Yes	63 (96.9%)	2 (3.1%)	0 (0.0%)	3.0 (0.2)	2.3 (1.2)	1.8 (1.4)
110020	Yes	86 (93.5%)	6 (6.5%)	1 (1.1%)	2.9 (0.3)	1.0 (0.0)	1.0 (0.0)
110021	Yes	83 (97.6%)	2 (2.4%)	0 (0.0%)	3.0 (0.2)	1.1 (0.2)	1.0 (0.0)

Figure S3: Subjective and objective presence of ON/OFF motor fluctuations. The percentage of non-ON beeps and OFF-beeps is insufficient to make valid comparisons on an individual level in most of the participants. Furthermore, there are large differences between the reported ON- and OFF-states and the experienced OFF-moments over days between patients who self-reported to have fluctuations. This prohibits On- vs Off-state analyses on individual levels and only allows us to make an analysis on group level, which cannot be used to draw hard conclusions.

References

1 Jankovic, J. Motor fluctuations and dyskinesias in Parkinson's disease: clinical manifestations. Movement disorders : official journal of the Movement Disorder Society 20 Suppl 11, S11-16, doi:10.1002/mds.20458 (2005).

2 Verhagen Metman, L. Recognition and treatment of response fluctuations in Parkinson's disease: review article. Amino acids 23, 141-145, doi:10.1007/s00726-001-0119-1 (2002).

3 Kim, A. et al. Emergence of non-motor fluctuations with reference to motor fluctuations in Parkinson's disease. Parkinsonism & related disorders 54, 79-83, doi:10.1016/j.parkreldis.2018.04.020 (2018).

4 van der Velden, R. M. J., Broen, M. P. G., Kuijf, M. L. & Leentjens, A. F. G. Frequency of mood and anxiety fluctuations in Parkinson's disease patients with motor fluctuations: A systematic review. Movement disorders : official journal of the Movement Disorder Society 33, 1521-1527, doi:10.1002/mds.27465 (2018).

5 Goetz, C. G. et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement disorders : official journal of the Movement Disorder Society 23, 2129-2170, doi:10.1002/mds.22340 (2008).

6 Hagell, P. & Nygren, C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. Journal of neurology, neurosurgery, and psychiatry 78, 1191-1198, doi:10.1136/jnnp.2006.111161 (2007).

7 Sanchez-Ferro, A. et al. New methods for the assessment of Parkinson's disease (2005 to 2015): A systematic review. Movement disorders : official journal of the Movement Disorder Society 31, 1283-1292, doi:10.1002/mds.26723 (2016).

8 Dorsey, E. R. et al. Moving Parkinson care to the home. Movement disorders : official journal of the Movement Disorder Society 31, 1258-1262, doi:10.1002/mds.26744 (2016).

9 Schneider, R. B. & Biglan, K. M. The promise of telemedicine for chronic neurological disorders: the example of Parkinson's disease. The Lancet. Neurology 16, 541-551, doi:10.1016/s1474-4422(17)30167-9 (2017).

10 Cancela, J. et al. Monitoring of motor and non-motor symptoms of Parkinson's disease through a mHealth platform. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2016, 663-666, doi:10.1109/EMBC.2016.7590789 (2016).

11 Tsiouris, K. M. et al. PD_Manager: an mHealth platform for Parkinson's disease patient management. Healthc Technol Lett 4, 102-108, doi:10.1049/htl.2017.0007 (2017).

12 Ferreira, J. J. et al. Quantitative home-based assessment of Parkinson's symptoms: the SENSE-PARK feasibility and usability study. BMC Neurol 15, 89, doi:10.1186/s12883-015-0343-z (2015).

13 Winberg, C. et al. The Use of Apps for Health in Persons with Multiple Sclerosis, Parkinson's Disease and Stroke - Barriers and Facilitators. Stud Health Technol Inform 242, 638-641 (2017).

14 Linares-Del Rey, M., Vela-Desojo, L. & Cano-de la Cuerda, R. Mobile phone applications in Parkinson's disease: A systematic review. Neurologia (Barcelona, Spain), doi:10.1016/j.nrl.2017.03.006 (2017).

15 Zhan, A. et al. Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity: The Mobile Parkinson Disease Score. JAMA neurology, doi:10.1001/jamaneurol.2018.0809 (2018).

16 Broen, M. P. et al. Unraveling the Relationship between Motor Symptoms, Affective States and Contextual Factors in Parkinson's Disease: A Feasibility Study of the Experience Sampling Method. PloS one 11, e0151195, doi:10.1371/journal.pone.0151195 (2016).

17 Lipsmeier, F. 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 33, 1287-1297, doi:10.1002/mds.27376 (2018).

18 Rodriguez-Molinero, A. et al. A Kinematic Sensor and Algorithm to Detect Motor Fluctuations in Parkinson Disease: Validation Study Under Real Conditions of Use. JMIR rehabilitation and assistive technologies 5, e8, doi:10.2196/rehab.8335 (2018).

19 Del Din, S., Godfrey, A., Mazzà, C., Lord, S. & Rochester, L. Free-living monitoring of Parkinson's disease: Lessons from the field. Movement Disorders 31, 1293-1313, doi:10.1002/mds.26718 (2016).

20 Vizcarra, J. A. et al. The Parkinson's disease e-diary: Developing a clinical and research tool for the digital age. Movement disorders : official journal of the Movement Disorder Society, doi:10.1002/mds.27673 (2019).

21 Espay, A. J. et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Movement disorders : official journal of the Movement Disorder Society 34, 657-663, doi:10.1002/mds.27671 (2019).

22 Verhagen, S. J., Hasmi, L., Drukker, M., van Os, J. & Delespaul, P. A. Use of the experience sampling method in the context of clinical trials. Evidence-based mental health 19, 86-89, doi:10.1136/ebmental-2016-102418 (2016).

23 van der Velden, R. M. J., Mulders, A. E. P., Drukker, M., Kuijf, M. L. & Leentjens, A. F. G. Network analysis of symptoms in a Parkinson patient using experience sampling data: An n = 1 study. Movement disorders : official journal of the Movement Disorder Society 33, 1938-1944, doi:10.1002/mds.93 (2018).

24 Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society 53, 695-699, doi:10.1111/j.1532-5415.2005.53221.x (2005).

25 PsyMate. www.psymate.eu

26 Heijmans, M. et al. Monitoring Parkinson's disease symptoms during daily life: a feasibility study. NPJ Parkinson's disease 5, 21, doi:10.1038/s41531-019-0093-5 (2019).

27 Palmier-Claus, J. E. et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta psychiatrica Scandinavica 123, 12-20, doi:10.1111/j.1600-0447.2010.01596.x (2011).

28 Ginty, A. T. in Encyclopedia of Behavioral Medicine (eds Marc D. Gellman & J. Rick Turner) 487-487 (Springer New York, 2013).

29 Simons, C. J. et al. Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. European psychiatry : the journal of the Association of European Psychiatrists 30, 900-906, doi:10.1016/j.eurpsy.2015.09.004 (2015).

30 Delespaul, P. A. E. G. Assessing schizophrenia in daily life : the experience sampling method. Doctor of Philosophy thesis, Maastricht University, (1995).

31 Ramdhani, R. A., Khojandi, A., Shylo, O. & Kopell, B. H. Optimizing Clinical Assessments in Parkinson's Disease Through the Use of Wearable Sensors and Data Driven Modeling. Frontiers in computational neuroscience 12, 72, doi:10.3389/fncom.2018.00072 (2018).

32 Thorp, J. E., Adamczyk, P. G., Ploeg, H. L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. Front Neurol 9, 1036, doi:10.3389/fneur.2018.01036 (2018).

33 Papapetropoulos, S. S. Patient diaries as a clinical endpoint in Parkinson's disease clinical trials. CNS neuroscience & therapeutics 18, 380-387, doi:10.1111/j.1755-5949.2011.00253.x (2012).

34 Janssens, K. A. M., Bos, E. H., Rosmalen, J. G. M., Wichers, M. C. & Riese, H. A qualitative approach to guide choices for designing a diary study. BMC Med Res Methodol 18, 140, doi:10.1186/s12874-018-0579-6 (2018).

35 Jacobs, N. et al. Deconstructing the familiality of variability in momentary negative and positive affect. Acta psychiatrica Scandinavica 127, 318-327, doi:10.1111/j.1600-0447.2012.01924.x (2013).

36 Hoff, J. I., van Hilten, B. J. & Roos, R. A. A review of the assessment of dyskinesias. Movement disorders : official journal of the Movement Disorder Society 14, 737-743 (1999).

37 Serrano, J. A. et al. Participatory design in Parkinson's research with focus on the symptomatic domains to be measured. Journal of Parkinson's disease 5, 187-196, doi:10.3233/JPD-140472 (2015).

38 Ferreira, J. J. et al. Clinical Parameters and Tools for Home-Based Assessment of Parkinson's Disease: Results from a Delphi study. Journal of Parkinson's disease 5, 281-290, doi:10.3233/JPD-140493 (2015).

Chapter 7

A long-term, real-life Parkinson monitoring database combining unscripted objective and subjective recordings

Habets JGV, Heijmans M, Leentjens AFG, Simons CJP, Temel Y, Kuijf ML, Kubben PL, Herff C

Published in Data 2021, 6(2)
Doi: https://doi.org/10.3390/data6020022

Abstract: Accurate real-life monitoring of motor and non-motor symptoms is a challenge in Parkinson's disease (PD). The unobtrusive capturing of symptoms and their naturalistic fluctuations within or between days can improve evaluation and titration of therapy. First-generation commercial PD motion sensors are promising to augment clinical decision-making in general neurological consultation, but concerns remain regarding their short-term validity, and long-term real-life usability. In addition, tools monitoring real-life subjective experiences of motor and non-motor symptoms are lacking. The dataset presented in this paper constitutes a combination of objective kinematic data and subjective experiential data, recorded parallel to each other in a naturalistic, long-term real-life setting. The objective data consists of accelerometer and gyroscope data, and the subjective data consists of data from ecological momentary assessments. Twenty PD patients were monitored without daily life restrictions for fourteen consecutive days. The two types of data can be used to address hypotheses on naturalistic motor and/or non-motor symptomatology in PD.

1. Summary

Parkinson's disease (PD)'s world-wide prevalence is expected to double to over 12 million patients by 2040 [1]. Current treatment strategies are symptomatic, mainly focus on improving motor function, and start with oral dopaminergic replacement medication. Refractory motor symptoms, adverse effects, or (non-)motor fluctuations can indicate advanced treatments, such as continuous levodopa administration or deep brain stimulation, or additional non-dopaminergic therapies [2,3]. A challenge in PD care is to improve symptom monitoring during patients' real life, in between clinical visits. Continuous, passive PD monitoring is suggested to improve therapy evaluation and titration by decreasing reliability on patient recall, and limitations of current monitor tools such as the lack of unobtrusive, repetitive assessment [4,5]. Wearable sensors monitoring motor symptoms are probably the best-known example of passive monitoring [6-9]. Wearable motion sensors are also suggested to investigate non-motor symptoms such as depression [10]. Additionally or complementary to continuous objective monitoring, continuous subjective monitoring via electronic (e-)diaries is suggested to contribute to both motor and non-motor PD monitoring in real life [11,12].

Due to the ubiquitous presence of smartphones and smartwatches with dedicated mHealth-applications, the collection and analysis of real-life data has increased exponentially over the last decade. Most of these devices allow objective motion data collection via inertial measurement units (IMUs). IMUs typically contain accelerometers and/or gyroscopes. In addition to these objective measures, subjective information about the patient's status can be obtained through diary methods or regular questionnaires administered via mHealth-applications. A valid translation of these real-life, or naturalistic, data into clinically or scientifically relevant information is an important challenge for researchers involved in PD and many (neuro)psychological and somatic diseases [6,12].

The presented multi-modal data were collected to improve understanding of the feasibility, usability, and validity of real-life PD monitoring, with a focus on motor symptoms. Objective data were collected via bilateral wrist and a chest IMU containing accelerometers and gyroscopes. Subjective data were collected via smartphone-based ecological momentary assessments (EMA), also called experience sampling methods. The practical feasibility of this novel combined method in a general PD population was demonstrated before [11]. We showed a good completion of objective and subjective data collection, with an acceptable burden for PD patients, and without high variability in completion between or within days.

These naturalistic long-term objective and subjective data aim to overcome the lack of unobtrusive, momentary, repetitive assessments of currently available PD motor monitoring devices [7,13,14]. Substantial concerns exist about the first-generation PD monitor devices regarding real-life validity, and their specific intended role in clinical practice [15,16]. Defining the exact role in clinical practice of a clinically supportive, data-driven tool is of critical importance to realize successful and impactful implementation [17,18]. In our data, the subjective data complements the motor monitoring by the objective data. Combining continuous objective and subjective data can help to overcome the well-known challenge of translating scripted, lab-based monitor methods, to unscripted, real-life monitor methods [19,20]. More specifically, it can serve as an alternative, continuous gold standard informing about subjectively experienced PD symptomatology, parallel to naturalistic sensor data. Traditional PD monitor instruments, such as the Movement Disorders Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) and

the Parkinson Disease Quality of Life Questionnaire (PDQ-39) are limited as they require the physical presence of trained clinicians, one assessment covers days to weeks, and their usability for longitudinal follow-up has been questioned [21-23]. As a proof-of-concept, we successfully predicted subjective EMA-answers regarding tremor severity based on objective motion data in a single participant [24].

Non-motor symptoms are notoriously disregarded in PD management, despite their repeatedly demonstrated effect on patients' quality of life [25]. Dedicated non-motor symptom scales are increasingly available and applied; however, naturalistic momentary assessment is no standard practice yet. EMA is explored for psychiatric disorders such as depression [26], and evidence suggests a role of EMA in monitoring or treating depression [27]. EMA may contribute to clinical practice and scientific research regarding non-motor symptoms in PD [28]; this hypothesis requires, however, further investigation.

This paper and its accompanying database aim to enable and stimulate PD researchers to answer research questions regarding real-life motor and/or non-motor symptoms, using objective data, subjective data, or both. For the patients who reported to suffer from motor fluctuations, motor-related motion sensor analyses can be performed. The potential value of these data is underlined by previous naturalistic PD monitoring work calling for continuous data sets containing self-reported motor states and motion sensor data [29]. Especially, the repetitive subjective symptom experiences via EMA offer new possibilities in motor monitoring development. Where recent validation studies aim on performance over longer time periods [9,30], the additional ESM facilitates analyses over shorter time windows. Moreover, the ESM offers a higher symptom resolution than merely ON-versus OFF-medication. Open-code efforts such as the BEAT-PD data challenge offer well-performing publicly available algorithms to differentiate between ON and OFF, and detecting tremors and dyskinesia [31]. These proposed signal processing and prediction analyses can be applied on finer-grain symptom resolutions in our dataset.

Further, many non-motor symptom-related research questions and hypotheses can be explored with these data. Non-motor symptoms are increasingly suggested to be part of continuous PD monitoring [12], and a concrete example is the suggestion to include patient-reported outcomes to monitor non-motor symptoms after deep brain stimulation for PD [32]. Research into general non-motor monitoring, as well as research into such specific non-motor monitoring applications, can profit from the data presented here.

With this data and this accompanying descriptor, we contribute to open-science for PD in general, and open-source, reproducible algorithms for real-life PD monitoring [29,31,33,34]. Furthermore, the applied methodology to combine momentary objective data with high-frequency objective data can be extrapolated to (EMA-)research in general.

2. Data Description

The data is publicly available via DataverseNL repository "EMA and wearable sensor monitoring in PD", under CC0 "Public Domain Dedication" license (https://doi.org/10.34894/5HHK8H) [35]. We provide raw sensor data and raw EMA data separately and unmerged, to enable every researcher to process the data as desired.

The code used for extracting, aligning, and merging both data types, including the example analysis, is available on https://github.com/jgvhabets/sensor_EMA_PD_monitoring.

2.1 Objective Sensor Data

Sensor data are stored in European Data Format [36] (edf)-files, and organized on the DataVerseNL repository in separate patient folders. Each patient folder contains multiple edf-files representing all 14 recording days.

- The name of every edf-file contains first the sensor name, followed by the start date and time of the recording. For example, in folder "110001", the file "13792_20180828_0752223.edf" contains the recording from participant 110001 with sensor 13792, which started recording on 28 August 2018, at 07:52:23. The read-me file "READ_ME_EMA_SENSOR_PD.txt" explains which sensor numbers represent left wrist, right wrist, or chest IMUs. The sensors actively recorded when they were not connected to a USB-charging device.
- Each edf-file contains six channels (representing the x-, y-, and z-axes for, respectively, accelerometer and gyroscope), including timestamps. Acceleration is recorded in m/s per second, and rotation is recorded in degrees per second.
- Prior to the first recording day, the clocks of all three sensors were reset and synchronized. The manufacturer assures temporal drift to be negligible over the period of two weeks with respect to merging and pairing with EMA-assessments.
- Single edf-files were created when a sensor was disconnected from the charger. A file continued storing data until the sensor was connected to a charger again and the file closed. If a recording passed midnight (00:00:00), the file closed as well, and a new file was created and continued storing data.

2.2. Subjective EMA Data

2.2.1. EMA Data Organization

The EMA method consisted of three types of questionnaires: a "beep" questionnaire, a daily morning questionnaire, and a daily evening questionnaire (see Chapter 6, Figure 1). The "beep" questionnaire was identically offered seven times per day, with an accompanying notification; the morning questionnaire was available to complete on own initiative between 06:00 and 12:00; and the evening questionnaire was available to complete on own initiative between 20:00 and 03:00.

The EMA data from all patients are stored in "EMA_data.csv".

The fist column provides a patient number, corresponding to the sensor data folder names.

Then, two columns provide timestamps indicating the start time and end time of beep-questionnaire completion.

These are followed by columns providing the answers on the items from the beep-questionnaire.

Then, columns provide the answers on the morning and evening questionnaires from the corresponding day.

The file "EMA_data_coding.xlsx" provides a clear explanation of the coding of all questionnaire items and answers.

2.2.2. EMA Content

If no multiple-choice is provided in square brackets, items are answered on a Likert scale from 1 to 7.

3. Methods

3.1. Participants

Twenty idiopathic PD patients participated in this study (see Chapter 6, Table 1). Patients were recruited via their treating neurologist or neurosurgeon at Maastricht University Medical Centre. Inclusion criteria were diagnoses of idiopathic PD, age between 18 and 80 years, in possession of a smartphone (minimal iOS 8 or Android 4), mastering spoken and written Dutch language, and available during two consecutive weeks of representative daily activities (meaning no holidays or planned hospital admission). No participants were excluded because of cognitive deficits (less than 24 points on the Montreal Cognitive Assessment). Hoehn and Yahr scores and levodopa equivalent daily dosages were collected. Twelve out of twenty participants (110002, 110003, 110004, 110006, 110008, 110014, 110016, 110017, 110018, 110019, 110020, and 110021) reported to suffer motor fluctuations despite dopaminergic (and deep brain stimulation) therapy. Disease-specific details are not included in the publicly available data set due to local ethical privacy regulations. Researchers can contact the authors when these data are requested. This study was approved by the local medical ethical committee and written informed consent was obtained from all participants, including the approval to share the anonymized data (METC azM/UM 2017-0307).

Participants were monitored via wearable sensors and EMA for 14 consecutive days and were instructed to not adjust their daily life routines or activities (Figure 1). An introductory meeting was held prior to day 1 at the patient's home during which the sensors were provided and the EMA application was installed and explained. A phone call consultation was performed at day 2 and day 8 to evaluate the progress and answer possible questions. After day 14, a researcher collected the sensors and evaluated the study period.

3.3. Parkinson's-Specific EMA Method

EMA is a validated method for observational studies or therapy evaluation in psychiatric and psychological diseases [37-39]. Subjective experiences are collected through questionnaires several times daily on semi-randomized moments. Recall-bias is minimized by allowing the patient to complete the questionnaire only during a short time window after notification, and by asking the patient to report his/her experience "at this moment". These repeated measurements aim to capture symptom fluctuations over the day, as well as slower fluctuating trends over longer time spans. The practical feasibility of this method in these 20 PD patients is described earlier [11]. We applied a PD-specific EMA questionnaire in this dataset, containing affect, motor, and non-motor symptomatology, as well as contextual items. The morning questionnaire contains mainly items regarding sleep and fitness after waking, and the evening questionnaire covers mainly motor functioning and symptomatology during the day, to enrich analyses on day levels. The development and internal validation of this PD EMA-questionnaire is extensively described earlier [40].

EMA and eDiaries are thought to be more sustainable tools to capture subjective intraday PD fluctuations in the future than paper diaries, which are more limited by procrastination, recall bias, and diary fatigue [41,42].

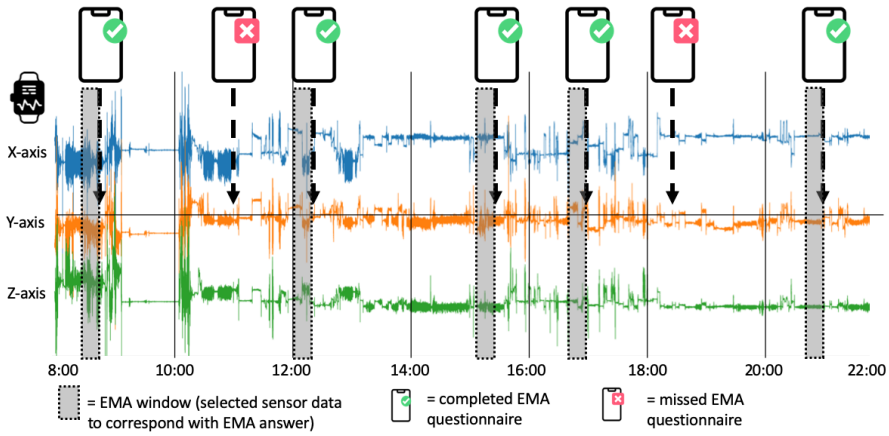


Figure 1. Schematic overview of one day of data collection. The blue, orange, and green lines represent, respectively, the signal on the x-, y-, and z-axis of an accelerometer, or a gyroscope. The signals are shown between 8:00 a.m. and 10:00 p.m. Selected time spans of sensor data which will be compared with the corresponding ecological momentary assessment (EMA) answers are shown in grey.

3.4. Devices

3.4.1. PsyMate (EMA Application)

The EMA monitoring was executed via a smartphone-based EMA application (PsyMate®, Maastricht, The Netherlands). The application presented the beep questionnaire seven times daily on semi-randomized moments, one questionnaire per two-hour window between 08:00 and 22:00 (see Figure 1). In order to reduce recall bias, the participants had to start answering the questionnaire within 15 min after notification. After these 15 min the questionnaire could not be opened anymore and was labeled as missed. The participants were instructed to complete as many questionnaires as possible without adjusting their lifestyle and activities. They were asked to complete the morning and evening questionnaire every day. These questionnaires were available during the above-mentioned time spans but were not presented with a notification.

The answers to completed beep questionnaires, as well as to the morning and evening questionnaires, were automatically uploaded to a server provided by the application developer. Missed questionnaires were not registered in the database.

3.4.2. MOX-5 (Wearable Sensor)

The patients used three wearable sensors, one located at each wrist and one at the sternum attached to a necklace (MOX5, Maastricht Instruments©, Maastricht, The Netherlands). The six degrees-of-freedom wearable sensors contained a tri-axial accelerometer and a tri-axial gyroscope, and recorded unprocessed raw data. The axial orientation of the wrist sensors was the following: x was parallel to the arm length, y recorded sideways movement in anatomical

position, and z recorded front/back movement in anatomical position. The axial orientation of the chest sensor was the following: x was parallel to the body length, y recorded movement sideways, and z recorded movement to the front or back. The accelerometer covered an amplitude range of ±8 g and the gyroscope covered a range of ±2000 degrees/s. Data were collected with a sampling rate of 200 Hz. The sensors stored all data on a built-in memory disk and did not automatically transfer data or provide real-time assessments. The participants were instructed to wear the sensors from the moment they rose in the morning, until they went to bed at night (ideally at least between 08:00 until 22:00), except during showering or bathing.

4. User Notes

4.1. Software

We performed our data pre-processing and analysis in Jupyter Notebooks for Python (Python version 3.6, Project Jupyter ©, https://jupyter.org, revision fe7c2909). We used software packages: pyedflib, pandas (version 0.24.2) [43], Numpy (version 1.16.4) [44], and scikit-learn (version 0.21.2) [45].

4.2. Interpretation of Data Quantity and Quality

Detailed descriptions of these results are reported earlier [11,40]. Most of the participants experienced the data collection as not incriminating (17 out of 20.85%), and 90% (18 out of 20) did not adapt their daily activities. On average, the participants wore the wearable sensors 94% of the instructed wearing time, resulting in almost 15 h of sensor data collected daily. EMA completion rates for beep, morning, and evening questionnaires were, respectively, 79%, 97%, and 94%. No differences were seen in completion between different study days, or different daily moments. For three participants (110007, 110010, and 110015), more than 25% of the sensor data corresponding to the EMA beeps were missing due to practical data storage issues. Internal validity of the EMA answers is explored by correlating subjectively reported concepts as positive versus negative affect, motor symptom severity, motor functioning, and dopaminergic medication states with each other. Positive correlations between positive affect and motor functioning and less symptom severity were hypothesized. Additionally, dopaminergic off-medication states were hypothesized to correlate positively with motor symptom severity, and to correlate negatively with motor functioning and positive affect. The evening questionnaire on motor functioning and motor symptoms during the past day were hypothesized to correlate in the same way as described above. On a group level, this hypothesis was confirmed by correlations linking positive affect, with fewer motor symptoms, better motor functionality, and on-medication states, and vice versa. On an individual level, not enough fluctuations were captured for every individual participant to reproduce this. Seventy-six percent of EMA beeps were answered in dopaminergic on-state, 21.5% in transition state between on- and off-state, and 2.5% in off-state. Mild fluctuations are therefore more likely to be captured more often in this data than severe fluctuations. Due to the high overall completion rates, and in-person evaluations with the participants, we suggest the low number of off-states can be explained by a low prevalence of true off-states. For medication-state analysis, it therefore can be suggested to compare on-state versus non-on-state (transition on/off plus off-state).

4.3. Combined Data Processing and Analyzing: Practical Example of Dopaminergic Fluctuation Detection

As an example, we propose a pipeline to align and merge two data types with varying sample frequencies. In this process, we decided to select sensor data from the 15 min preceding the start of EMA beep questionnaire completion. These 15 min are chosen because the patients were instructed to complete the questionnaire according to their experiences at that exact moment ("How do you feel at this moment"). The 15 min block was a pragmatic and arbitrary decision; researchers might deviate from this approach based on their own arguments or hypothesis.
To support the usability of the data, we will give an example of dopaminergic fluctuation detection in participant 110018, suffering from strong symptom fluctuations [24]. We will merge both data types, extract features from the sensor data, and present correlation between the objective and subjective data by predicting the subjective reported medication state based on the objective sensor features. We extract features designed for wrist-sensor data, and will repeat classification analyses based on wrist- and chest-sensor data. The wrist-sensor analysis is comparable to an analysis in previous work [11]. Here, we repeat this analysis with data from the chest-sensor. Although the wrist-data and the chest-data were recorded during the same time points, we hypothesize that the classification analysis based on wrist-data is more successful to differentiate medication-state.

4.3.1. Data Merging
To enable merging of the EMA data and the sensor data, the completed beep questionnaires from the EMA data were used as a reference frame. For sensor data extraction, the timestamps of all starting moments of completed beep questionnaires were extracted and the raw sensor data corresponding to each 15-min block preceding these EMA time stamps were selected. Raw sensor data were first down-sampled to 100 Hz since a higher sample rate was not necessary for the intended analysis. This resulted in a data frame containing 90,000 rows of sensor data (15 min of 100 Hz) corresponding to every complete beep questionnaire. Each row consisted of three values (x-, y-, z-axis) for each of the three accelerometers and the gyroscopes.

4.3.2. Sensor Data Pre-Processing and Feature Extraction
Features representing the research topic of interest, in this case, dopaminergic medication fluctuation, have to be extracted from the raw, 100 Hz sensor data. Depending on the features and hypothesis, researchers can be interested in feature values per, e.g., 60 s of extracted sensor data. This time span is referred to as the feature window length. We extracted various features designed for wrist-accelerometer and wrist-gyroscope data representing bradykinesia or medication fluctuation [8]. By calculating features, a new dataset is created, in which each row contained the timestamp and the feature value for one feature window length.
For this practical example provided, we applied a feature window length of 900 s [24], and did not apply overlapping epochs in the sensor data. The following temporal domain features were extracted: the root mean square (RMS) of every accelerometer-axis feature window length, and the amplitude range, representing the range between the minimal and maximum amplitude per accelerometer-axis feature window length. The following spectral domain features were extracted: the spectral power density for frequencies corresponding to tremors (3.5–7.5 Hz) for both accelerometer-axes and gyroscope-axes, and the spectral power density for frequencies corresponding to bradykinesia (0.5–3 Hz) per accelerometer-axis, the dominant frequency per

accelerometer-axis, and the dominant energy ratio per accelerometer-axis (dividing the energy within the dominant frequency by the total sum of the energy in all frequencies).

4.3.3. Classification Analysis

The corresponding EMA items on medication status were used as binary labels for the extracted sensor data. We trained a logistic regression model in a 5-fold cross-validation. Applied on wrist-worn sensor data, the classifier differentiated between the medication states with an area under the receiver operator curve (AUROC) of 0.73 (Figure 2A). Applied on chest-worn sensor data, the same classifier was not able to discriminate between the medication states above chance level (AUROC of 0.51) (Figure 2B).

The presented examples show that subjectively reported experiences of medication fluctuation can be predicted from features derived from the motion sensor data.

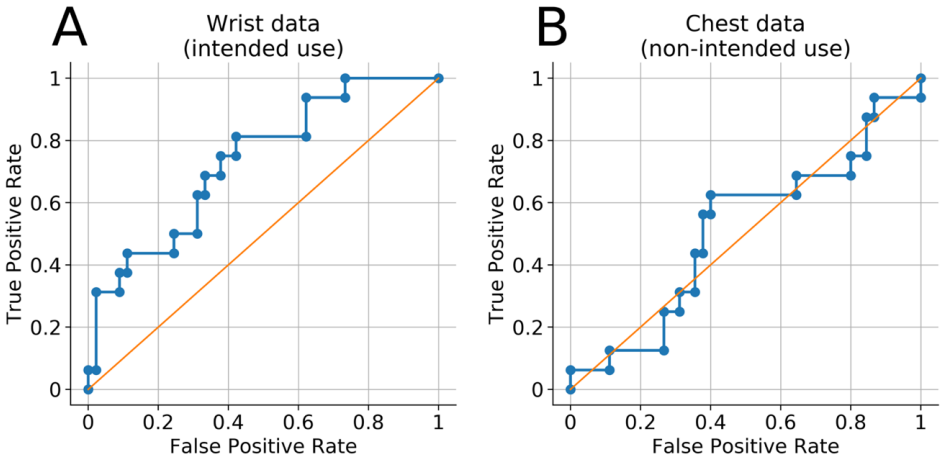


Figure 2. Receiver Operator Curves (ROC) for differentiating medication-state of a patient with symptom fluctuations based on sensor features which are developed for wrist-worn sensor data. (**A**) ROC for analysis with features extracted from wrist-worn data. (**B**) ROC for analysis with features extracted from chest-worn data. Figure 2A is based on comparable analyses as in our feasibility study [11].

Acknowledgments: We want to thank the ESM Expert Group of Maastricht University/ Open University Heerlen for their advice on the design of the EMA methodology. We want to thank Karel Borkelmans for the development and technical support on the PsyMate application. We want to thank the clinical multidisciplinary Deep Brain Stimulation/ Movement Disorders team of Maastricht UMC+ for advice during the design of the study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dorsey, E.R.; Sherer, T.; Okun, M.S.; Bloem, B.R. The Emerging Evidence of the Parkinson Pandemic. Journal of Parkinson's disease 2018, 8, S3-s8, doi:10.3233/jpd-181474.
2. Odin, P.; Ray Chaudhuri, K.; Slevin, J.T.; Volkmann, J.; Dietrichs, E.; Martinez-Martin, P.; Krauss, J.K.; Henriksen, T.; Katzenschlager, R.; Antonini, A., et al. Collective physician perspectives on non-oral medication approaches for the management of clinically relevant unresolved issues in Parkinson's disease: Consensus from an international survey and discussion program. Parkinsonism & related disorders 2015, 21, 1133–1144, doi:10.1016/j.parkreldis.2015.07.020.
3. Seppi, K.; Weintraub, D.; Coelho, M.; Perez-Lloret, S.; Fox, S.H.; Katzenschlager, R.; Hametner, E.M.; Poewe, W.; Rascol, O.; Goetz, C.G., et al. The Movement Disorder Society Evidence-Based Medicine Review Update: Treatments for the non-motor symptoms of Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2011, 26 Suppl 3, S42-80, doi:10.1002/mds.23884.
4. Papapetropoulos, S.; Mitsi, G.; Espay, A.J. Digital Health Revolution: Is it Time for Affordable Remote Monitoring for Parkinson's Disease? Frontiers in Neurology 2015, 6, doi:10.3389/fneur.2015.00034.
5. 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, doi:10.3233/jpd-181498.
6. Odin, P.; Chaudhuri, K.R.; Volkmann, J.; Antonini, A.; Storch, A.; Dietrichs, E.; Pirtosek, Z.; Henriksen, T.; Horne, M.; Devos, D., et al. Viewpoint and practical recommendations from a movement disorder specialist panel on objective measurement in the clinical management of Parkinson's disease. NPJ Parkinson's disease 2018, 4, 14, doi:10.1038/s41531-018-0051-7.
7. Pahwa, R.; Isaacson, S.H.; Torres-Russotto, D.; Nahab, F.B.; Lynch, P.M.; Kotschet, K.E. Role of the Personal KinetiGraph in the routine clinical assessment of Parkinson's disease: recommendations from an expert panel. Expert review of neurotherapeutics 2018, 18, 669–680, doi:10.1080/14737175.2018.1503948.
8. Thorp, J.E.; Adamczyk, P.G.; Ploeg, H.L.; Pickett, K.A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. Front Neurol 2018, 9, 1036, doi:10.3389/fneur.2018.01036.
9. Powers, R.; Etezadi-Amoli, M.; Arnold, E.M.; Kianian, S.; Mance, I.; Gibiansky, M.; Trietsch, D.; Alvarado, A.S.; Kretlow, J.D.; Herrington, T.M., et al. Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease. Science Translational Medicine 2021, 13, eabd7865, doi:10.1126/scitranslmed.abd7865.
10. Choi, K.W.; Chen, C.Y.; Stein, M.B.; Klimentidis, Y.C.; Wang, M.J.; Koenen, K.C.; Smoller, J.W. Assessment of Bidirectional Relationships Between Physical Activity and Depression Among Adults: A 2-Sample Mendelian Randomization Study. JAMA Psychiatry 2019, 76, 399–408, doi:10.1001/jamapsychiatry.2018.4175.
11. Heijmans, M.; Habets, J.G.V.; Herff, C.; Aarts, J.; Stevens, A.; Kuijf, M.L.; Kubben, P.L. Monitoring Parkinson's disease symptoms during daily life: a feasibility study. NPJ Parkinson's disease 2019, 5, 21, doi:10.1038/s41531-019-0093-5.
12. Vizcarra, J.A.; Sanchez-Ferro, A.; Maetzler, W.; Marsili, L.; Zavala, L.; Lang, A.E.; Martinez-Martin, P.; Mestre, T.A.; Reilmann, R.; Hausdorff, J.M., et al. The Parkinson's disease e-diary: Developing a clinical and research tool for the digital age. Movement disorders : official journal of the Movement Disorder Society 2019, 10.1002/mds.27673, doi:10.1002/mds.27673.
13. Neurotechnologies, G.L. Kinesia ProViewTM. Availabe online: (accessed on 20-11-2017).
14. Rodriguez-Molinero, A.; Perez-Lopez, C.; Sama, A.; de Mingo, E.; Rodriguez-Martin, D.; Hernandez-Vara, J.; Bayes, A.; Moral, A.; Alvarez, R.; Perez-Martinez, D.A., et al. A Kinematic Sensor and Algorithm to Detect Motor Fluctuations in Parkinson Disease: Validation Study Under Real Conditions of Use. JMIR rehabilitation and assistive technologies 2018, 5, e8, doi:10.2196/rehab.8335.
15. Warmerdam, E.; Hausdorff, J.M.; Atrsaei, A.; Zhou, Y.; Mirelman, A.; Aminian, K.; Espay, A.J.; Hansen, C.; Evers, L.J.; Keller, A. Long-term unsupervised mobility assessment in movement disorders. The Lancet Neurology 2020, 19, 462–470.
16. Fasano, A.; Mancini, M. Wearable-based mobility monitoring: the long road ahead. The Lancet. Neurology 2020, 19, 378–379.
17. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019, 17, 195, doi:10.1186/s12916-019-1426-2.
18. Pencina, M.J.; Goldstein, B.A.; D'Agostino, R.B. Prediction Models - Development, Evaluation, and Clinical Application. The New England journal of medicine 2020, 382, 1583–1586, doi:10.1056/NEJMp2000589.
19. Kluge, F.; Del Din, S.; Cereatti, A.; Gaßner, H.; Hansen, C.; Helbostadt, J.L.; Klucken, J.; Küderle, A.; Müller, A.; Rochester, L., et al. Consensus based framework for digital mobility monitoring. medRxiv 2020, 10.1101/2020.12.18.20248404, 2020.2012.2018.20248404, doi:10.1101/2020.12.18.20248404.
20. Galperin, I.; Hillel, I.; Del Din, S.; Bekkers, E.M.J.; Nieuwboer, A.; Abbruzzese, G.; Avanzino, L.; Nieuwhof, F.; Bloem, B.R.; Rochester, L., et al. Associations between daily-living physical activity and laboratory-based assessments of motor severity in patients with falls and Parkinson's disease. Parkinsonism & related disorders 2019, 62, 85–90, doi:10.1016/j.parkreldis.2019.01.022.
21. Goetz, C.G.; Tilley, B.C.; Shaftman, S.R.; Stebbins, G.T.; Fahn, S.; Martinez-Martin, P.; Poewe, W.; Sampaio, C.; Stern, M.B.; Dodel, R., 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, doi:10.1002/mds.22340.
22. Hagell, P.; Nygren, C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. Journal of neurology, neurosurgery, and psychiatry 2007, 78, 1191–1198, doi:10.1136/jnnp.2006.111161.

23. Evers, L.J.W.; Krijthe, J.H.; Meinders, M.J.; Bloem, B.R.; Heskes, T.M. 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, doi:10.1002/mds.27790.
24. Heijmans, M.; Habets, J.; Kuijf, M.; Kubben, P.; Herff, C. Evaluation of Parkinson's Disease at Home: Predicting Tremor from Wearable Sensors. In Proceedings of 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 23-27 July 2019; pp. 584-587.
25. Todorova, A.; Jenner, P.; Ray Chaudhuri, K. Non-motor Parkinson's: integral to motor Parkinson's, yet often neglected. Practical Neurology 2014, 14, 310, doi:10.1136/practneurol-2013-000741.
26. Seppala, J.; De Vita, I.; Jamsa, T.; Miettunen, J.; Isohanni, M.; Rubinstein, K.; Feldman, Y.; Grasa, E.; Corripio, I.; Berdun, J., et al. Mobile Phone and Wearable Sensor-Based mHealth Approaches for Psychiatric Disorders and Symptoms: Systematic Review. JMIR Ment Health 2019, 6, e9819, doi:10.2196/mental.9819.
27. Simons, C.J.; Hartmann, J.A.; Kramer, I.; Menne-Lothmann, C.; Hohn, P.; van Bemmel, A.L.; Myin-Germeys, I.; Delespaul, P.; van Os, J.; Wichers, M. Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. European psychiatry : the journal of the Association of European Psychiatrists 2015, 30, 900–906, doi:10.1016/j.eurpsy.2015.09.004.
28. Broen, M.P.; Marsman, V.A.; Kuijf, M.L.; Van Oostenbrugge, R.J.; van Os, J.; Leentjens, A.F. Unraveling the Relationship between Motor Symptoms, Affective States and Contextual Factors in Parkinson's Disease: A Feasibility Study of the Experience Sampling Method. PloS one 2016, 11, e0151195, doi:10.1371/journal.pone.0151195.
29. Mahadevan, N.; Demanuele, C.; Zhang, H.; Volfson, D.; Ho, B.; Erb, M.K.; Patel, S. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. NPJ Digit Med 2020, 3, 5, doi:10.1038/s41746-019-0217-7.
30. Rodriguez-Molinero, A. Monitoring of Mobility of Parkinson's Patients for Therapeutic Purposes - Clinical Trial (MoMoPa-EC). Availabe online: (accessed on 18 November).
31. MJFF, S. BEAT-PD DREAM Challenge (by Sage Bionetworks; Michael J. Fox Foundation). Availabe online: (accessed on January 7).
32. Loring, D.W.; Block, C.; Staikova, E.; Miocinovic, S. Patient-Reported Outcomes Measurement Information System (PROMIS) Assessment of Non-Motor Features in Deep Brain Stimulation Candidates: Relationship to the Beck Depression and Anxiety Inventories. Archives of Clinical Neuropsychology 2020, 10.1093/arclin/acaa091, doi:10.1093/arclin/acaa091.
33. Bloem, B.R.; Marks, W.J., Jr.; Silva de Lima, A.L.; Kuijf, M.L.; van Laar, T.; Jacobs, B.P.F.; Verbeek, M.M.; Helmich, R.C.; van de Warrenburg, B.P.; Evers, L.J.W., et al. The Personalized Parkinson Project: examining disease progression through broad biomarkers in early Parkinson's disease. BMC Neurol 2019, 19, 160, doi:10.1186/s12883-019-1394-3.
34. Rochester, L.; Mazzà, C.; Mueller, A.; Caulfield, B.; McCarthy, M.; Becker, C.; Miller, R.; Piraino, P.; Viceconti, M.; Dartee, W.P., et al. A Roadmap to Inform Development, Validation and Approval of Digital Mobility Outcomes: The Mobilise-D Approach. Digital biomarkers 2020, 4, 13–27, doi:10.1159/000512513.
35. Habets, J.; Kubben, P. EMA and wearable sensor monitoring in PD. V2 ed.; Maastricht University, M., Ed. DataverseNL: 2020; doi:10.34894/5HHK8H.
36. Kemp, B.; Varri, A.; Rosa, A.C.; Nielsen, K.D.; Gade, J. A simple format for exchange of digitized polygraphic recordings. Electroencephalogr Clin Neurophysiol 1992, 82, 391–393, doi:10.1016/0013-4694(92)90009-7.
37. Verhagen, S.J.; Hasmi, L.; Drukker, M.; van Os, J.; Delespaul, P.A. Use of the experience sampling method in the context of clinical trials. Evidence-based mental health 2016, 19, 86–89, doi:10.1136/ebmental-2016-102418.
38. Brannon, E.E.; Cushing, C.C.; Crick, C.J.; Mitchell, T.B. The promise of wearable sensors and ecological momentary assessment measures for dynamical systems modeling in adolescents: a feasibility and acceptability study. Translational behavioral medicine 2016, 6, 558–565, doi:10.1007/s13142-016-0442-4.
39. Palmier-Claus, J.E.; Myin-Germeys, I.; Barkus, E.; Bentley, L.; Udachina, A.; Delespaul, P.A.; Lewis, S.W.; Dunn, G. Experience sampling research in individuals with mental illness: reflections and guidance. Acta psychiatrica Scandinavica 2011, 123, 12–20, doi:10.1111/j.1600-0447.2010.01596.x.
40. Habets, J.; Heijmans, M.; Herff, C.; Simons, C.; Leentjens, A.F.; Temel, Y.; Kuijf, M.; Kubben, P. Mobile Health Daily Life Monitoring for Parkinson Disease: Development and Validation of Ecological Momentary Assessments. JMIR Mhealth Uhealth 2020, 8, e15628, doi:10.2196/15628.
41. Papapetropoulos, S.S. Patient diaries as a clinical endpoint in Parkinson's disease clinical trials. CNS neuroscience & therapeutics 2012, 18, 380–387, doi:10.1111/j.1755-5949.2011.00253.x.
42. Erb, M.K.; Karlin, D.R.; Ho, B.K.; Thomas, K.C.; Parisi, F.; Vergara-Diaz, G.P.; Daneault, J.F.; Wacnik, P.W.; Zhang, H.; Kangarloo, T., et al. mHealth and wearable technology should replace motor diaries to track motor fluctuations in Parkinson's disease. NPJ Digit Med 2020, 3, 6, doi:10.1038/s41746-019-0214-x.
43. Jeff Reback, W.M., jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, … Martin Winkel Pandas 1.1.1, 2020.
44. Walt, S.v.d.; Colbert, S.C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 2011, 13, 22–30, doi:10.1109/MCSE.2011.37.
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. Journal of machine learning research 2011, 12, 2825–2830.

Chapter 8

Evaluation of Parkinson's Disease at Home: Predicting Tremor from Wearable Sensors

Heijmans M* & Habets JGV*,
Kuijf ML, Kubben PL, Herff C
* Authors contributed equally

Published as conference paper on the Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2019 (584-487)
Doi: https://doi.org/10.1109/EMBC.2019.8857717

Abstract

The continuous monitoring of Parkinsons's disease (PD) symptoms would allow to automatically adjust medication or deep brain stimulation parameters to a patient's momentary condition. Wearable sensors have been proposed to monitor PD symptoms and have been validated in a number of lab and hospital settings. However, taking these sensors into the daily life of patients introduces a number of difficulties, most notably the absence of an observable ground truth of what the user is currently doing.

In this pilot study, we investigate PD symptoms by combining wearable sensors on both wrist and the chest with a questionnaire based evaluation of PD symptoms, in the form of experience sampling method. For a tremor dominant patient, we show that experienced tremor severity can be predicted
from the sensor data with correlations of up to r = 0:43. We evaluated different window lengths to calculate the features in and see better results for longer window lengths. Our results show that continuous monitoring of PD symptoms in daily life is feasible using wearable sensors.

Introduction

Parkinson's disease (PD) is a degenerative disorder in which a loss of dopaminergic neurons in the substantia nigra causes motor and non-motor symptoms. Cardinal motor symptoms are bradykinesia, rigidity, tremor and postural instability. The vast majority of PD patients is treated with dopaminergic medication, e.g. levodopa, to restore dopamine levels in the basal ganglia. A well-known challenge of this pharmacological treatment is the occurrence of intra-daily motor fluctuations and levodopa-induced dyskinesia after five to ten years [1]. These fluctuations are also known as ON- and OFF-fluctuations, where ON-state refers to periods during which motor symptoms are well treated, and OFF-state refers to periods during which motor symptoms are not treated sufficiently. Suffering from ON-OFF fluctuations can be an indication for deep brain stimulation (DBS), which then aims to increase the time a patient spends in ON-state. Therefore, monitoring these intra-daily fluctuations is of importance to control and adjust dopaminergic restoration therapy, and maybe even DBS in the future [2].

Wearable sensors, in particular inertial measuring units consisting of accelerometers and/or gyroscopes, have potential to monitor PD symptoms and ON-OFF fluctuations in a continuous, non-obtrusive manner. Further validation of systems and preferably care-models is needed before they are applicable for daily life monitoring [3,4]. Studies monitoring tremor during daily life activities are often limited because data were collected in simulated home settings in either a research lab or hospital [5–9]. Although the real-life representation of a simulated home setting stays debatable, videotape recordings of the assessments can be made and be used as parallel 'ground truth' data. Studies which are performed in the real home situation are however limited because of the lack of such parallel ground truth data [3]. This highlights one of the major challenges in PD monitoring; the availability of reliable ground truth data on the patients clinical state.

We introduce experience sampling method (ESM), also known as ecological momentary assessment, to provide additional subjective data on the patient's well-being to verify and evaluate our wearable sensor data [10]. The smartphone-based ESM method we used, presents digital questionnaires at semi-randomized moments throughout the day asking the patient about current motor state and symptoms. Additionally, there are accompanying questions on mood, affect and context, which can also fluctuate and cannot be measured directly from the wearable sensors. The development and validation of the ESM-method will be described elsewhere. In this case report, we describe a tremor-dominant PD patient who reported to suffer from ON-OFF fluctuations. We demonstrate how fluctuations in tremor severity, measured by ESM, can be predicted from wearable sensors during daily living.

Material and methods

Study Design

During a period of two consecutive weeks, participants had to wear three wearable sensors (one at each wrist and one at the chest) and had to complete ESM questionnaires.

1) Wearables: This study used MOX5 wearables (Maastricht Instruments, Maastricht, The Netherlands), containing both an accelerometer and gyroscope. The accelerometer covered an amplitude range of ±8 g and the gyroscope covered a range of ±2000 deg/s. Data were collected with a sampling rate of 200 Hz. The wearables were attached to the body via handmade wristbands and a necklace (Fig. 1). Participants had to wear the wearables only during daytime, and had to charge them at night. For subsequent alignment to the ESM questionnaires, the data were time stamped on the devices. Data were saved on the device and were extracted by the research team after the measurement period.

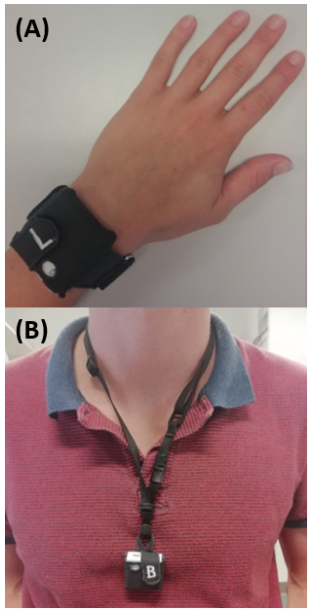


Figure 1. Wrist-worn (A) and chest (B) sensors including accelerometer and gyroscope as well as flash storage and batteries. Participants wear the sensors from waking up until they go to bed and charge them during the night.

2) ESM: The ESM app Psymate was installed on the participants smartphone. We developed a specific PD questionnaire using previous work [11,12] and by patient and clinician interviews about what is considered important when identifying inter- and intra-daily fluctuations in PD symptoms. During the measurement period, participants received repetitive questionnaires (containing 26 questions) at seven semi-randomized moments during the day. These questionnaires contained questions on mood, affect, context, motor state and PD symptoms at that specific moment (e.g. I experience tremor). Participants had to rate the questions on a 7-point Likert scale. The questionnaires stayed available for only 15 minutes, thereby excluding recall bias. The participants were asked to complete as much questionnaires as possible without adapting their normal daily behaviour.

Participants

This study was approved by the METC azM/UM and written informed consent was obtained from all participants. Recruitment was done by neurologists, neurosurgeons, and PD nurses at the Maastricht University Medical Centre. Inclusion criteria were: Diagnosed with idiopathic PD; in possession of a smartphone (iOS or android); mastering Dutch language; and being available for two consecutive weeks of representative daily activities. Patients were excluded if they scored less than 24 points on the Montreal Cognitive Assessment [13]. We included and finished data collection in 20 participants.

To be sure to have sufficient fluctuations in tremor severity for this pilot study, we investigate the case of a participant that reported to be tremor dominant and to suffer from severe ON-OFF fluctuations. This participant is a 65-year old male who has been diagnosed with idiopathic PD for 6 years. He scored 26 points on the Montreal Cognitive Assessment. Since one year this participant receives bilateral subthalamic DBS therapy. In addition to the DBS therapy, he takes dopaminergic medication five times a day (levodopa equivalent daily dose = 788). The patient reported ON-OFF fluctuations which were medication related, resulting in periods of increased

tremor ranging from approximately half an hour before standard medication intake moments until a quarter after medication intake.

Data processing

In order to align the continuous sensor data to the relevant subjective evaluations measured through ESM, we extracted the 15 minutes of sensor data prior to each completed questionnaire. We hypothesize that the previous 15 minutes best reflect the participants symptomatic experience. Each
of these 15 minute long blocks is then associated with the corresponding answers from the ESM questionnaire. To extract meaningful information from the sensor data, we extracted features both in the time and spectral domain [14]. We evaluate different window lengths for the feature extraction. Fig. 2 visualizes our feature extraction procedure.

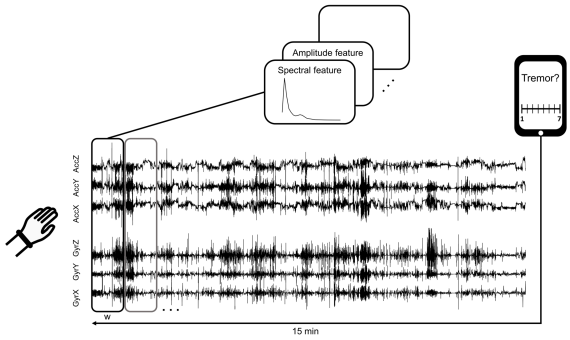


Fig. 2. Feature extraction procedure. For each ESM questionnaire, we extract the proceeding 15 minutes of continuous sensor data. We divide these into windows of varying length w and extract different features. The extracted feature vectors are then annotated with the answers from the corresponding questionnaire.

Feature extraction

Based on previous work [7,9,15] in the detection and decoding of tremor and bradykinesia in lab or hospital settings, we extracted the following features:

1) Logarithmic signal energy in the 3.5-7.5 Hz frequency band. We extracted this information from both accelerometer and gyroscope data. (6 features)
2) Total signal energy in the form of the root mean squared signal. This feature was extracted from the low-pass filtered (3 Hz) accelerometer data. (3 features)
3) The dominant frequency in the low-pass filtered (3 Hz) accelerometer data. (3 features)
4) The dominant energy ratio, which we calculated by dividing the maximum energy by the total energy in the low-pass filtered (3 Hz) accelerometer data. (3 features)
5) The amplitude range, which we extracted from the lowpass filtered (3 Hz) time series of accelerometer data. (3 features)
6) Maximum normalized cross-correlation and corresponding temporal offset among all pairs of low-pass filtered accelerometer time-series data. (2 features)

This results in a total of 20 features for each of the three wearables and thus a total of 60 features. To investigate the effect of window length and see which window length best represents the patient's answers, we extracted these features in non-overlapping windows of

30, 60, 120, 180, 300 and 900 seconds. We assigned the same answers from the corresponding questionnaire to all windows extracted from one block of sensor data. Please note that this leads to different amounts of samples for the different window length conditions. While the 900 second long windows result in 62 samples (= number of filled-out questionnaires), the 30 second long windows result in 62 * 30 = 1860 samples. This also results in vastly different baseline levels, which we estimate using permutation levels (see Section Data Analysis).

Data Analysis

For this pilot study, we focused on both wrist worn sensors and excluded the chest sensor. This results in a total feature space of 40. To test the prediction of tremor severity from the extracted features, we applied a 10-fold cross-validation. In this approach, 90% of the data are used for training and the remaining 10% are used for testing, this is repeated until all samples have been used for testing exactly once. To predict the continuous severity assessments, we employed a simple linear regression. We calculated Pearson correlations between original and predicted tremor severity to evaluate the prediction.

We tested statistical significance through random permutation tests, in which we randomly shuffled the severity scores and calculated correlations between original and shuffled scores. This process was repeated 1000 times and the 95% highest correlations were used as a random baseline to signify the α = 0:05-level.

Results

Correlation coefficients between original and predicted tremor scores are significantly above chance level for all window lengths except for the 120 second long windows (Fig. 3 (A)). This drop is further characterized by higher correlation coefficients for shorter (above 0.15 for both 30 and 60 seconds) and longer windows (180 seconds and above). Highest correlation scores were obtained for 15 minute long windows with r = 0:43. Fig. 3 (B) shows an example of predicted and actual tremor scores for 15 minute long windows.

(A) (B)

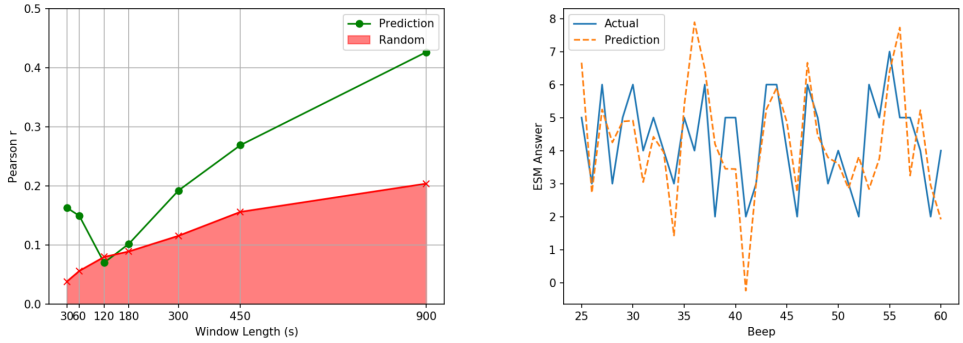


Fig. 3. Mean correlation results for tremor severity (A). Correlation coefficients for all but the 120 second window length are significantly above chance level (red shaded area, alpha = 0:05). Example of actual (blue) and predicted (yellow) tremor scores for the full 15 minute window condition (B).

Discussion

The evaluation of PD symptoms in continuous monitoring during daily life activities is very challenging, as no ground truth is available. Patients might sit still in front of the TV, be engaged in a conversation or take a walk in the park. These different activities will have tremendous effects on the measured sensor data and variations need to be taken into account before reliable prediction of symptoms can be achieved. Our preliminary results give compelling evidence that features that are known to be successful in the prediction of PD symptoms in lab settings [15], can be used to predict tremor severity during the patients daily life. These results expand previous work which did make use of activity detection using videotapes, clinical assessments, and predefined motor tasks [6–8,8,9]. In addition, our results expand previous work which predicted clinical tremor scores from tremor data which was recorded while patients were comfortably seated in a chair [16]. Lastly, this study is to our knowledge the first one comparing the effect of using different window lengths for tremor feature extraction. Longer window lengths resulted in better correlation results of the predicted ESM scores. It is important to note that the chance level results are also higher for longer windows, as the number of samples decreases, and high correlations are more likely in smaller sample sizes.

Clearly, our results need to be extended to more patients with different amounts of ON-OFF fluctuations and to other PD motor symptoms, such as bradykinesia, rigidity, and dyskinesia.

Conclusion

In this case report, investigating fluctuating PD tremor in a patient's daily life, we show that subjective tremor scores can be predicted with good correlations from wrist-worn sensors using very simple regression models. We employ standard features from the literature and investigate the influence of different window lengths on the prediction quality. For analyses of this individual patient, longer window lengths seem to result in better prediction quality. This confirms the possibility and increases the knowledge on monitoring PD symptoms in a daily life situation.

References

1. Jankovic, J. Motor fluctuations and dyskinesias in Parkinson's disease: clinical manifestations. *Mov Disord* **20 Suppl 11**, S11-16 (2005).
2. Habets, J. G. V. *et al.* An update on adaptive deep brain stimulation in Parkinson's disease. *Mov Disord* **33**, 1834–1843 (2018).
3. Rodríguez-Molinero, A. *et al.* A Kinematic Sensor and Algorithm to Detect Motor Fluctuations in Parkinson Disease: Validation Study Under Real Conditions of Use. *JMIR Rehabil Assist Technol* **5**, (2018).
4. Thorp, J. E., Adamczyk, P. G., Ploeg, H.-L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. *Front Neurol* **9**, 1036 (2018).
5. Cole, B. T., Roy, S. H., De Luca, C. J. & Nawab, S. Dynamic neural network detection of tremor and dyskinesia from wearable sensor data. *Annu Int Conf IEEE Eng Med Biol Soc* **2010**, 6062–6065 (2010).
6. Cole, B. T., Roy, S. H., De Luca, C. J. & Nawab, S. H. Dynamical learning and tracking of tremor and dyskinesia from wearable sensors. *IEEE Trans Neural Syst Rehabil Eng* **22**, 982–991 (2014).
7. Hoff, J. I., Wagemans, E. A. & van Hilten, B. J. Ambulatory objective assessment of tremor in Parkinson's disease. *Clin Neuropharmacol* **24**, 280–283 (2001).
8. Roy, S. H., Cole, B. T., Gilmore, L. D., De Luca, C. J. & Nawab, S. H. Resolving signal complexities for ambulatory monitoring of motor function in Parkinson's disease. *Annu Int Conf IEEE Eng Med Biol Soc* **2011**, 4832–4835 (2011).
9. Salarian, A. *et al.* Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system. *IEEE Trans Biomed Eng* **54**, 313–322 (2007).
10. Brannon, E. E., Cushing, C. C., Crick, C. J. & Mitchell, T. B. The promise of wearable sensors and ecological momentary assessment measures for dynamical systems modeling in adolescents: a feasibility and acceptability study. *Transl Behav Med* **6**, 558–565 (2016).
11. Ferreira, J. J. *et al.* Clinical Parameters and Tools for Home-Based Assessment of Parkinson's Disease: Results from a Delphi study. *J Parkinsons Dis* **5**, 281–290 (2015).
12. Serrano, J. A. *et al.* Participatory design in Parkinson's research with focus on the symptomatic domains to be measured. *J Parkinsons Dis* **5**, 187–196 (2015).
13. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* **53**, 695–699 (2005).
14. Herff, C. & Krusienski, D. J. Extracting Features from Time Series. in *Fundamentals of Clinical Data Science* (eds. Kubben, P., Dumontier, M. & Dekker, A.) (Springer, 2019).
15. Patel, S. *et al.* Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. *IEEE Trans Inf Technol Biomed* **13**, 864–873 (2009).
16. Jeon, H. *et al.* Automatic Classification of Tremor Severity in Parkinson's Disease Using a Wearable Device. *Sensors (Basel)* **17**, (2017).

Chapter 9

Rapid dynamic naturalistic monitoring of bradykinesia in Parkinson's disease using a wrist-worn accelerometer

Habets JGV, Herff C, Kubben PL, Kuijf ML, Temel Y, Evers LJW, Bloem BR, Starr PA, Gilron R* & Little SJ*
* Authors contributed equally

In preparation, May 2021

Abstract

Introduction: Motor fluctuations in Parkinson's disease are defined as inconsistent therapeutic benefits on symptoms such as bradykinesia and rigidity and impair the quality of life of many Parkinson patients. Traditional Parkinson evaluation tools are not designed for continuous, naturalistic symptom monitoring. Objective motor fluctuation monitoring should accurately measure and decode movement, during real world (naturalistic) activities. Although the use of commercially available motor fluctuation monitoring devices for multiple days can augment neurological decision making, the feasibility of rapid and dynamic decoding of fine-grained motor fluctuations using wearable accelerometry is unclear. We investigate the performance of machine learning classification models identifying rapid (on a minute level), medication-induced wearing-off motor fluctuations.

Methods: As part of the Parkinson@Home study protocol, Parkinson patients were encouraged twice to perform an hour of unconstrained activities in their own homes, in deprived and optimal dopaminergic medication states, while being recorded with bilateral wrist-accelerometers. Naturalistic bradykinesia-representing features were extracted from unilateral accelerometer data of 20 patients, from the bodyside with the largest unilateral bradykinesia fluctuation. After comparing the accelerometer features on the hour level, medication-state classification analyses were performed on the minute-level. The influence of individual versus group training data, window length, and amount of training data was analyzed.

Results: Statistically significant classification of medication induced bradykinesia fluctuations were seen in 90% of the Parkinson patients at the single minute timescale using either the group or individual model. Individually trained models performed equally despite the small training dataset and unconstrained motor activities. Bradykinesia classification of the group models improved as the length of the feature windows was widened up to 300 seconds, and as the number of training patient datasets was increased.

Conclusion: Rapid, naturalistic Parkinson motor monitoring has important clinical potential to evaluate dynamic symptomatic and therapeutic fluctuations. Rapid short-term Parkinson monitoring is subject to different conceptual challenges than longer-term Parkinson monitoring, which should be addressed in separate model development and validation studies.

Introduction

Parkinson's disease (PD) is a disabling neurodegenerative disorder characterized by motor and non-motor symptoms that significantly affect patients' motor performance and quality of life (QoL) [1–3]. Symptomatic PD management initially focuses on pharmacological dopamine replacement therapies [4]. However, half of PD patients develop 'wearing-off' motor fluctuations during the first decade after diagnosis [5,6]. Wearing-off motor fluctuations are defined as inconsistent therapeutic benefits on symptoms such as bradykinesia and rigidity, despite regular dopaminergic delivery [6]. These motor fluctuations and also other dopaminergic related side effects can markedly impair patients' QoL [7]. Motor fluctuations are therefore a primary indication for consideration of deep brain stimulation (DBS) [1,8]. Adequate monitoring of motor fluctuations is essential for treatment evaluation, both in the presence and absence of DBS, and wearable motion sensing represents an appealing approach to support this [9,10], although several challenges remain to be addressed [11,12].

Ideally, objective motor fluctuation monitoring should accurately measure and decode movement, during real world (naturalistic) activities, and be simple to implement for patients [10,13]. Currently used Parkinson evaluation tools such as the Movement Disorders Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) and the Parkinson Disease QoL questionnaire (PDQ-39) are not designed for chronic continuous, naturalistic symptom monitoring [14,15]. They contain questionnaires which capture subjective estimates of retrospective symptoms over a week (MDS-UPDRS II and IV), or a month (PDQ-39), but these are dependent on patient recall, which is often in perfect, particularly in patients with cognitive decline. Observing and scoring motor fluctuations requires trained health providers to perform single time point evaluations (MDS-UPDRS III). Motor diaries, often used as gold standard for 24-hour naturalistic monitoring, require self-reporting every 30 minutes [16]. This burden causes recall-bias and diary fatigue in long-term use [17].

The strong clinical need for continuous symptom tracking, together with the wide availability of affordable accelerometer sensors via the ubiquitous presence of digital devices, has led to numerous academic and commercially available wearable PD monitoring systems [18–22]. So far, promising augmentation of clinical decision making has been reported, based on commercially available motor fluctuation monitoring devices that were used for multiple days, prior to neurological consultation [21,23–26]. However, there is evidence that motor fluctuation estimates from wearable PD monitoring systems correlate better with longer term clinical metrics over time windows of days rather than hours [21,27]. There is a notable difference between these longer time windows, yielding good correlations with clinical gold standards or being applied in validation protocols, and the smaller time windows being described in development studies [19,21,28–30]. Whether rapid and dynamic motor fluctuation decoding using wearable accelerometer data can be accurate enough to serve fine grained precision medicine needed for dynamic optimization of closed loop therapy systems remains to be determined.

Previously reported motor decoding systems have been typically trained on group data. However, using these group level classification models, personalized motion sensor algorithms have been suggested to have the potential to improve individualized PD management and monitoring [20,31].

Here, we investigate the performance of machine learning classification models identifying rapid (on a minute level), medication-induced wearing-off motor fluctuations in PD patients. The

classification models are trained on unconstrained naturalistic (at home) motion data derived from a wrist-worn accelerometer. Classification models based on individual data are compared with models based on group data. Before these predictive classification models are tested, long term statistical differences are investigated between the pre- and post-medication recordings at the group level. Further, we analyse the influence of the amount of training data, and the length of analysed accelerometer data epochs (time window lengths), on classification results. We focused symptom decoding on bradykinesia since this cardinal feature of PD [1,32] has been found to be more challenging to detect with motion sensors than tremor or dyskinesia, likely due to higher distributional kinematic overlap with normal movements [18,33–35].

We hypothesize that individualized motion classification models in PD would demonstrate more reliable short-term classification of naturalistic bradykinesia fluctuations compared to group models.

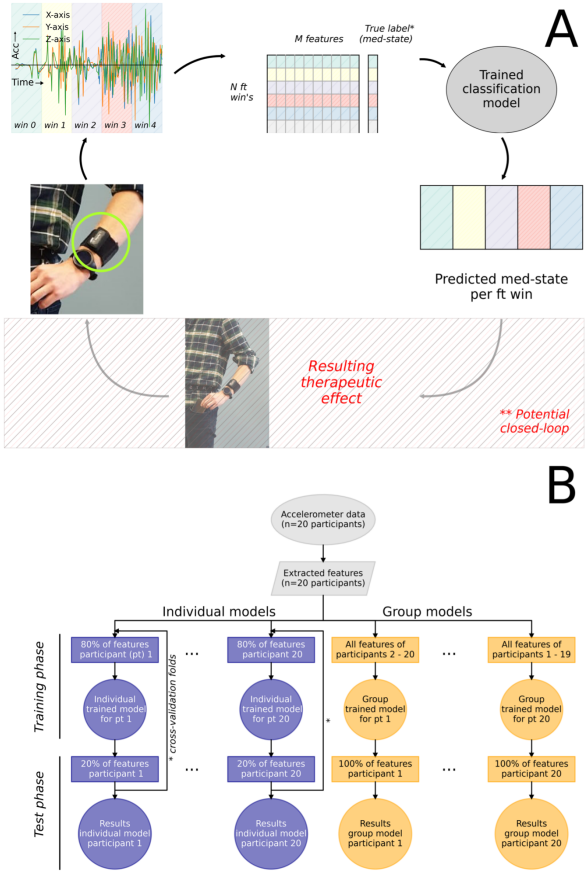


Figure 1: Motion based Parkinsonian motor fluctuation detection workflows.
A: From center left: Bilaterally wrist motion-sensors (Gait Up Physilog 4, Gait Up SA, CH) are worn by the patient (green circled) and continuously collected tri-axial accelerometer data (top left) from the body side with the severest bradykinesia fluctuation are analyzed. For each 'feature window' (win 0, win 1, etc), a number of (M) features from the temporal and spectral domain (Table S1) are extracted from the raw accelerometry (acc) data. This results in a data set of M columns (features) and N rows (N feature windows), with each row providing all M feature values calculated per feature window (win) (top center). In the training phases, the true labels (*) are included in the analysis and represent the true medication states. In the test phases, only the feature values are included, the true labels are unknown, and the previously trained model (top right) predicts a medication state for each feature window (center right) which can then be compared with the true labels to determine the performance. The shaded bottom area (**) represents a potential therapeutic closed-loop, in which the generated predictions result in therapy changes, and influence the present clinical state of the patient (center left).
B: Workflow to train and test individual and group models. Identical features were extracted from the raw accelerometer data (grey symbols) for every individual

participant. For the individually trained models (blue), the features from 80% of a participant's epochs were used in the *training phase* (y-axis). The trained individual model was tested with the remaining, unused, 20% of epochs during the *test phase*. The arrows (*) from test phase to training phase represent the multiple cross-validation folds applied to train and test the individual models on different selections of training and test data. For the group models (yellow), each participant was tested in turn, with data from the other 19 participants used in the *training phase*.

Results

Study population and recorded data

20 PD patients from the Parkinson@Home data repository [36] were included in this study. The Parkinson@Home study recorded accelerometer data in two medication-states, while PD patients were encouraged to perform an hour of their unconstrained activities in their own homes. MDS-UPDRS III scores were assessed right before the recordings by trained physicians directly in the patients' homes. The first MDS-UPDRS assessment took place in the dopaminergic deprived state (pre-medication), and the second assessment after patients experienced the full effect of their usual dopaminergic medication (post-medication). We included PD patients who showed an improvement (larger than 0) in the sum of unilateral MDS-UPDRS III items representing upper extremity bradykinesia, in at least one body side (see Methods section). Wrist-accelerometer data only from the side with the largest unilateral upper extremity bradykinesia improvement were included for feature extraction and classification analyses (see Figure 1).

Demographic and disease specific characteristics are presented in Table 1. In total, 3138 minutes of accelerometer data were recorded in the 20 included patients. After balancing the data sets for medication status, 2380 minutes of accelerometer data were included. On average 59.5 (+/- 14.3) minutes of accelerometer data from both pre- and post-medication recordings were included per participant. We extracted multiple features which are described in the current literature to represent bradykinesia in naturalistic wrist-accelerometer recordings (see Table S1 for details and references). In total, 103 motion accelerometer features were extracted for every feature window, describing both temporal and spectral domain characteristics.

Characteristics	
Total number (% female)	20 (60%)
Age (years, mean (sd))	63.4 (6.4)
Accelerometer data per medication state (minutes, mean (sd))	59.5 (14.3)
Accelerometer data per medication state, after activity filtering (minutes, mean (sd))	44.5 (13.9)
PD duration (years, mean (sd))	8.1 (3.5)
Levodopa equivalent daily dosage (milligrams, mean (sd))	959 (314)
MDS-UPDRS III pre-medication MDS-UPDRS III post-medication	43.8 (11.6) 27.1 (9.6)
MDS-UPDRS fluctuations* - Total score part III change - Hand bradykinesia - Body bradykinesia	 16.7 (8.6) 3.9 (2.0) 0.9 (0.6)

- Leg bradykinesia	0.9 (3.1)
- Hand Tremor	2.1 (2.3)
- Gait	1.0 (1.0)
- Posture	0.5 (1.1)
- Facial expression	1.2 (0.9)

*: fluctuation in (sub-)score between pre- and post-medication recording.

Table 1: Demographic and disease specific characteristics of patient population

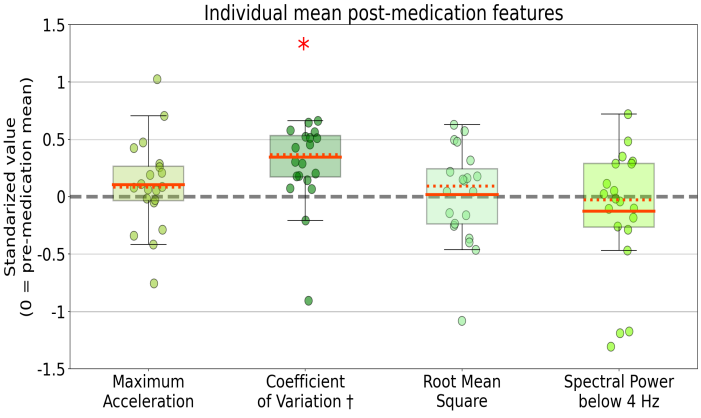


Figure 2: Distributions of individual means for four main movement features. Colored dots represent individual mean feature values during post-medication recording. Individual post-medication mean values are standardised using the individual pre-medication recordings as a reference. The red asterix indicates a significant difference on group level between mean coefficient of variations of pre- and post-medication means (alpha=0.05, MANOVA and post-hoc analysis, FDR corrected). † = one positive outlier (1.7) not visualized.

Group level statistical analysis of cardinal motion features across medication states

First, we compared pre- and post-medication accelerometer recordings at the group level using four accelerometer features which represent the commonly used motion features implemented in naturalistic bradykinesia signal processing (maximum acceleration magnitude, coefficient of variation of acceleration magnitude, root mean square of acceleration, and spectral power (below 4 Hz)) [18,37]. The individual mean values per whole medication-state recording were compared at the group level.

The pre- and post-medication recordings significantly differed at the group level based on the individual mean values of the four main accelerometer-features (MANOVA, Wilk's lamba = 0.389, F-value = 14.2, $p < 0.001$). Post-hoc repeated measures ANOVAs demonstrated that only the individual coefficient of variation averages significantly differed between pre- and post-medication states ($p = 0.0042$) (figure 2). The coefficient of variation (calculated from the absolute tri-axial signal vector) holds the strongest potential to distinguish medication states using longer term accelerometer data at the group level.

Machine learning classification of short window medication states

Next, to test classification performance over short time windows (60 second accelerometer feature windows), support vector (SV) and random forest (RF) machine learning models were applied. To tailor the classification models to different activity levels, we repeated all analysis including an activity filter (see Methods section). First, the same four cardinal features were included in analysis, and thereafter the full feature set. The medication state classification based

on the previously selected four motion features (using SV and RF models, trained with group and individual data) led in most participants to low AUC scores (means per model ranged between 0.49 and 0.64) and low accuracies (means per model ranged between 49% and 60%) (see table S2 for detailed results).
Classification analyses based on the full feature sets (103 features), led to higher AUC scores and classification accuracies per model, and per participant (table S2, and figures S3AB), regardless of SV or RF model, or trained with group or individual data. Mean AUC scores per model ranged between 0.65 and 0.70, and mean accuracies per model ranged between 60% and 65% (table S2). Most participants yielded AUC scores and accuracies significantly better than chance level (17 out of 20 participants per model), tested through random surrogate dataset generation.

The best individual and group models together, classified medication states per 60 seconds significantly better than chance level in 90% of participants (18 out of 20) (figure 3 and table S2). In 12 participants AUC scores were statistically significantly higher than classification based on our surrogate dataset for both individual and group models, 5 only for group models, and 1 only for the individual model (figure 3). Both individual and group models resulted in mean AUC scores of 0.70 (+/- respectively 0.18 and 0.10), and mean accuracies of respectively 65% (+/- 0.14) and 64% (+/- 0.08) over all 20 participants (figure 3, table S2). A trend of additional value of individualized model training could be seen at an individual level when the participants not exceeding chance levels were disregarded. Half of the participants (10 out of 20) benefited from individual model training with a higher AUC score than for their group model. Notably, the individual models resulted in several participants with AUC scores below chance level (figure 3).
These findings confirm the feasibility to differentiate between rapid, short-term, naturalistic medication states based on accelerometer recordings. No statistically significant difference was seen between models trained with individual and group data. To improve insights in which methods drive rapid, dynamic motor fluctuation classification, we analysed the applied methodologies in detail.

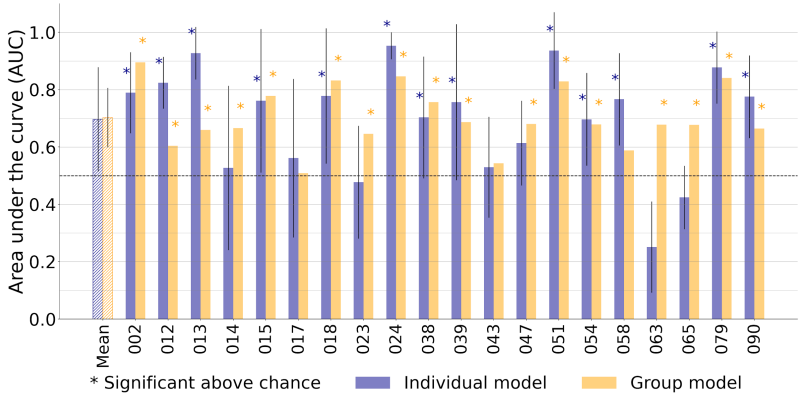


Figure 3: Classification of medication induced motor fluctuations on short accelerometer time windows in individual participants. The first pair of bars represents the mean area under the curve (AUC) score over the twenty participants. Each subsequent pair of bars (002 to 090) represents the AUC scores from one participant. The blue bars represent the AUC score for the individual model, and the yellow bars represent

the group model. Note that for the individual models, AUC scores are the averages over the multiple cross-validation folds within a participant (figure 1B). The asterisks indicate whether the corresponding AUC score was significantly better than chance level (5000-repetitions permutation test). Both models have equal mean AUC scores. It is notable that the majority (18 out of 20 of participants) has at least one significant score. Half of the participants yielded a higher AUC score with the individual model than with the group model.

Optimal methodological approaches for short window medication states classification

To understand which methodological approaches yielded best performance, we explored the differences between individual and group trained models and the effect of an activity filter in detail and investigated optimal training data sizes and feature window lengths (see Methods section below).

Due to the classification superiority of models (for 60 seconds feature windows) using 103 features (compared to 4 feature models), only models using 103 features are explicitly analysed going forward (AUC p-values < 0.002, accuracy p-values < 0.020, significance tested via a 5000-permutation test described in supplementary figure 3AB). SV individual models resulted in higher AUC scores and accuracies than RF individual models in 15 out of 20 patients (figure S3C, $p = 0.009$). SV and RF group models yielded similar AUC scores and accuracies (figure S3D, $p = 0.406$). Overall, applying the activity filter led to slightly better mean results per model (table S2). On an individual level, there was no significant difference between classification performance with or without activity filtering (figure S3EF, p-values ranged between 0.06 and 0.41). However, it was noted that there was a trend towards higher individual predictive performance with activity filtering.

Classification performance of bradykinesia focussed motor fluctuations and co-occurring tremor

To ensure that the described predictive performance did not rely on co-occurring tremor fluctuations (despite the bandwidth-filtering of tremor frequencies; see Methods section), we explored the relation between SV classification performance and clinical bradykinesia and tremor fluctuation. At a group level, both bradykinesia and tremor fluctuations (unilateral upper extremity MDS-UPDRS sub items, see Methods) correlated not significantly, but weakly with AUC scores in the individually trained models (bradykinesia spearman $r = 0.24$ ($p = .305$), tremor spearman $r = 0.34$ ($p = .140$). The AUC scores for the group models did not show a relevant correlation with bradykinesia or tremor fluctuations (spearman r's respectively = 0.01 ($p = .962$) and 0.11 ($p = .642$)). Similar low correlations were found for the accuracy scores and the clinical fluctuations (table S3). At an individual level, we found significant AUC scores in participants with (13, 24, and 79) and without (39, 51, and 58) tremor fluctuations (figure S4). Individual predictive performance is found not to be proportional to the size of tremor fluctuations, which suggests feasibility for non-tremor dominant PD patients.

Influence of training data size and feature window length

We next sought to determine the number of patients needed to train a group level model with good classification performance. We found an increase in the predictive performance (AUC) of the group models as the number of patient datasets used during model training was increased (figure 4A). The mean AUC showed the steepest increase while increasing the amount of patients included in training data towards 10 subjects. Afterwards the increase in mean AUC flattened towards 19 included participants.

Next we also wanted to investigate the impact of the accelerometer data feature window length on the predictive performance of the group models. Increasing the length of the feature windows

up to 300 seconds improved the mean AUC (figure 4B). Due to data size limitations, the feature windows were not expanded further than 300 seconds. These analyses could not be reproduced for the individual models due to data size limitations.

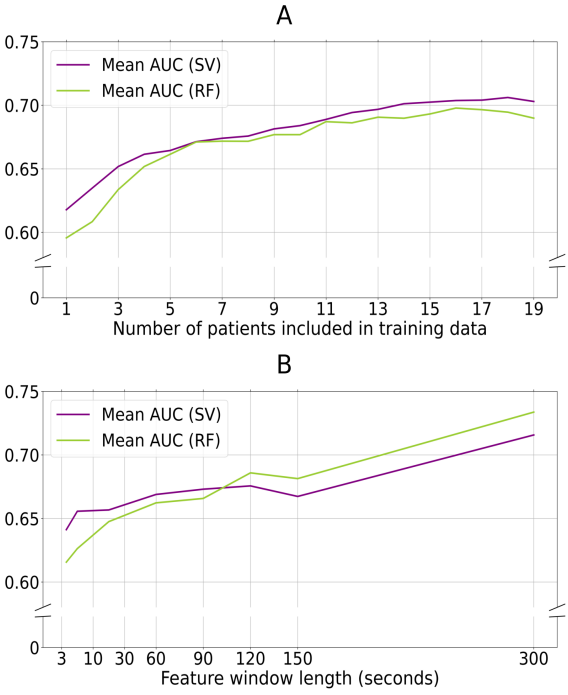


Figure 4: Increasing number of training patients and long window duration improves classification performance.
A: Group models are trained for every patient with a varying number of included training data, (x-axis). On the y-axis, the AUC is shown for both SV and RF models, both included activity filtering. An increase in AUC is seen for SV and RF models parallel to an increase in included training patients.
B: Group models are trained for every patient with various feature window lengths (x-axis). On the y-axis, the AUC of the SV and RF models are visualized. Due to the longer feature window lengths, we did not apply the activity filter in this sub-analysis to deal with data size limitations. Larger window lengths up to 300 seconds increased classification performance, while smaller window lengths decrease classification performance.
AUC: area under the receiver operator characteristic; SV: support vector classifier; RF: random forest classifier.

Discussion

Our results demonstrate that rapid dynamic classification of medication induced bradykinesia is possible in PD patients at the single minute timescale using a wrist worn accelerometer. Using a large number of accelerometer motion metrics aimed for naturalistic bradykinesia differentiation and machine learning, we demonstrated statistically significant classification of medication states could be achieved in 90% of participants (18 out of 20) using either the group model or the individual model (figure 3). Using only a single input feature we found a significantly increased coefficient of variation of accelerometer amplitude using longer time scales of data (~60 mins) (figure 2). The hypothesized additional value of individualized model training to classify short-term, naturalistic motor fluctuations was not found as both individual and group models resulted in a mean AUC of 0.70 (figure 3, table S2). Expansion of the duration of accelerometer features in the group trained classification models up to 300 seconds, as well as inclusion of more patients in the training data, may facilitate improvement of rapid dynamic motor fluctuation detection (figure 4).

Clinical relevance of rapid dynamic motor fluctuations detection

Naturalistic PD monitoring systems evaluated over sustained time windows have been proposed to augment neurological consultation [20,21,23] and improve pharmacological trial assessments [11,38]. Further suggested naturalistic PD monitor implementations include real-time feedback [22], such as medication adjustment or adaptive deep brain stimulation evaluation [39–41], or programming [42]. Implementation of wearable monitoring for long-term clinical evaluation versus short-term closed-loop therapy support will most likely require wearable monitor systems validated at different time scales. Therefore, separate development and validation studies are required per specific intended clinical utilization. By doing so, the questioned limits of temporal resolution in naturalistic PD motor monitoring will also be further clarified [10,12,43]. We here show that rapid classification of clinical bradykinesia states is possible using group level models with the potential to inform and benefit rapid, closed-loop, approaches to therapy adjustment.

The hypothesized additional value of individual model training

At a group level, individual model training did not lead to higher mean predictive performance. At an individual level, individual model training did lead to both higher and lower AUC scores (figure 3). This may be explained by the smaller size of training data in the individual trained models compared to the group trained models (figure 1B). These results neither indicate superiority, nor inferiority of individualized models with our current dataset. Therefore, replication of individually trained models using more data per subject is indicated. We predict that increasing training data included in individual model training will result in classification superiority of individual models. Currently it is not known at which amount of training data this will occur.

Increasing individual training data sets does not only increase the total amount of training, it also allows for the application of larger feature windows which will likely increase classification performance as well (figure 4) [34,44]. The expansion of the feature window length has to be weighed against the time window required for a desired real-time clinical application.

Challenges towards clinically impactful naturalistic PD motor monitoring

The here presented findings focussed, besides on individual data analysis, on the temporal scale of data analysis (feature window length), and on the temporal scale of generated outcome (aggregation of multiple output values towards clinically meaningful value). Both relate to a major challenge in naturalistic PD monitoring, which is choosing a gold standard for model development and validation. Since naturalistic PD monitoring aims to identify continuous fluctuations in symptom severity, a repetitive gold standard, feasible and valid in the naturalistic real-life situation, is needed. Ideally, it discriminates motor symptoms more detailed than the extreme pre- versus post-medication states. Scientists have tried to overcome this challenge by analysing the clinical effect on neurological decision making of long-term, multi-day monitoring [24,29], or by aggregating and generalizing data of multiple months into average day and week curves and comparing these with a clinician's overall impression of the patient's fluctuation pattern [21]. Also, PD specific eDiaries are suggested to collect momentary patient experiences which can be used as gold standard, after validation of the specific eDiary [45–48]. Currently, no ideal solution is available, and we support such creative solutions, as the short-term scores reasonably correlate with available (different time-scaled) gold standards, and the long-term application considers clinical utilization and impact [43]. Reason to not expect perfect correlations between novel PD monitor systems and established rating scales may be their potential superiority [10,13]. 'Gold standards are constantly challenged and superseded when appropriate' [49].

A second major challenge is to design and implement wearable PD monitor models with high predictive performance, without losing generalization to real world situations due to the use of highly constrained, artificial movement sets [11,50]. Our models' numerical predictive performance is at least comparable with the majority of available PD monitoring models focussed on bradykinesia or hypokinesia [18,22,33,35]. Models reporting higher predictive performance applied for example stricter gait detection [20], or collected data during constrained experimental protocols [18]. However, it should be noted that differences in temporal epochs of training and testing between our study (short duration) and other commercial devices (long duration) make direct comparison challenging [18–21]. Lower predictive performance and higher inter-individual variation in performance are known challenges of unconstrained activities during data collection [34,37]. This is explained by the higher variance in captured activities in training and test data, and the subsequent aggregation of different activity classes, such as walking and gross and fine motor tasks, that are known to perform differently in classification models [20,34]. An open-source, naturalistic activity classifier which standardizes naturalistic activity classification, can be a next step towards better symptom detection without losing real-world generalizability.

When a model is able to generate valid symptom severities (either short or long term), a next challenge of combining and interpreting these scores has to be overcome. First, to provide evidence that our models detected bradykinesia rather than simply labelled periods as bradykinetic due to presence of tremor signals, we explored the dependence on the relationship between our bradykinesia classification with tremor. At a group level, individual tremor and bradykinesia fluctuations correlated similarly with individual classification results (table S3). At an individual level, both participants with and without tremor fluctuations demonstrated good bradykinesia classification results (figure S4). This supports our hypothesis that bradykinesia related motor fluctuation detection is feasible independently from tremor. Next, future development and validation of naturalistic PD monitor systems should provide clear instructions for clinicians how to interpret and implement a model in clinical practice, rather than only reporting predictive performance. This translation, from (several) numerical output values to results or recommendations understandable for a clinician, will be essential to realize clinical impact in the future [43,51].

Future scientific opportunities to improve naturalistic PD monitoring development

First, we predict that the coming expansion of real-world motion data sets, containing long-term data over weeks to years in patients with PD, will support optimization of individually trained models [52]. These larger datasets will also allow the exploration of alternative, more data-dependent, computational analyses such as deep reinforcement learning and neural networks including long short-term memory [34,53]. Moreover, unsupervised models could also be explored to overcome the previously discussed gold standard dilemma by surpassing the need of long-term, repetitive, true labels [11,54]. For unsupervised model development, the potential of the coefficient of variation of wrist acceleration might be of value to differentiate bradykinetic motor fluctuations (figure 3). Interestingly, the coefficient of variation of neural beta signals have also been shown to be discriminative for motor state in PD [55].

Second, open-source research initiatives should catalyse the development of naturalistic PD monitor models which are not dependent on proprietary software [10,37]. The Mobilize-D

consortium for example, introduced a roadmap to standardize and structure naturalistic PD monitoring by creating specific 'unified digital mobility outcomes' [56,57]. During the development of these outcomes, features describing distribution ranges and extreme values, rather than means or medians, should be considered [11]. Parallel to open-source initiatives, other creative collaborations between industry and academia such as data-challenges might offer valuable (interdisciplinary) cross-fertilization [58].
Further, adding more limb sensors to improve naturalistic PD monitoring is controversial. Although good results are reported with combined wrist and ankle sensors [59], additional limb sensors are not felt likely to improve performance according to others [34,60,61]. The additional value of smart insoles for naturalistic PD monitoring is not yet described, and might play a role in the future [62].

Limitations
Our study was limited most by the individual data set sizes, which restricted inferences that could be made regarding models trained with individual versus group data. Also, individual differences in movement quality and quantity during the unconstrained pre- and post-medication recordings likely led to unbalanced datasets in terms of captured activities during the two medication states. This imbalance compromises pattern recognition based data analysis [34]. The latter limitation is inherent to naturalistic PD monitoring [18], and exploring the boundaries of this limitation is essential for future PD monitor applications. Replication of our methodologies in larger data sets, and inclusion of validated activity classifiers may contribute to overcoming this limitation.

Conclusion
We here demonstrate that naturalistic, short-term, bradykinesia fluctuation monitoring at the minute time scale is feasible in PD patients using a wrist-worn accelerometer and group or individually trained machine learning models. Bradykinesia classification improved with the increasing window length of accelerometer data tested and with the number of patients trained in the group model.
Individually trained models were equivalent to group trained models in terms of classification performance despite the small training dataset and unconstrained motor activities. This suggests that rapid, dynamic PD motor monitoring has important clinical potential for closed loop medication and therapy adjustments. Rapid short-term PD monitoring is subject to different conceptual challenges than longer-term PD monitoring, which should be addressed in model development and validation studies.

Methods

Study sample

We have used data from the Parkinson@Home validation study [36]. Detailed descriptions of the study's protocol and feasibility have been described earlier [63,64]. In brief, the study recruited 25 patients diagnosed with PD by a movement disorders neurologist who were all undergoing treatment with oral levodopa therapy. All participants had wearing-off motor fluctuations (MDS-UPDRS part IV item 4.3 ≥1) and at least slight Parkinson-related gait impairments (MDS-UPDRS part II item 2.12 ≥1 and/or item 2.13 ≥1). Participants who were treated with advanced therapies (DBS or infusion therapies) or who suffered significant psychiatric or cognitive impairments which hindered completion of the study protocol were excluded.

For the current subset of PD patients, we excluded three participants who did not show a levodopa-induced improvement in unilateral upper extremity bradykinesia, on both sides (equal or less than zero points). Unilateral upper extremity bradykinesia was defined as the sum of MDS-UPDRS part III items 3c, 4b, 5b, and 6b for the left side, and items 3b, 4a, 5a, and 6a for the right side. Sum scores from medication on-states were compared with sum scores from medication off-states. For each included participant, only data from the side with the largest clinical change in upper extremity bradykinesia sub items were included. Two participants were excluded because there was less than 40 minutes of accelerometer data available from their pre- or post-medication recording.

The study protocol was approved by the local medical ethics committee (Commissie Mensgebonden Onderzoek, region Arnhem-Nijmegen, file number 2016-1776). All participants received verbal and written information about the study protocol and signed a consent form prior to participation, in line with the Declaration of Helsinki. The de-identified dataset will be made available to the scientific community by the Michael J Fox Foundation.

Data were recorded via bilateral wrist-worn wearable devices (Gait Up Physilog 4, Gait Up SA, CH). For our current analysis, only unilateral tri-axial accelerometer data were analysed. Data recording was performed at the participants' homes. Recordings consisted of two sessions which took place on the same day. First, the pre-medication recording was performed in the morning after overnight withdrawal of dopaminergic medication. Second, the post-medication recording was performed when the participants experienced the full clinical effect after intake of their regular dopaminergic medication. During both recordings, participants performed an hour of unconstrained activities within and around their houses. At the start of both recordings, a formal MDS-UPDRS III was conducted by a trained clinician.

Data pre-processing and feature extraction

Accelerometer data were sampled at 200 Hz and down sampled to a uniform sampling rate of 120 Hertz (Hz) using piecewise cubic interpolation. The effect of gravity was removed from each of the three time series (x-, y-, and z-axes) separately, by applying a 'l1-trend filter' designed to analyse time-series with an underlying piecewise linear trend [65]. Time series were low-pass filtered at 3.5 Hz to attenuate frequencies typically associated with Parkinsonian tremor in accelerometer time series [66]. In addition to the three individual accelerometer time series, we computed a composite time series containing the vector magnitude of the three individual accelerometer axes $[x^2 + y^2 + z^2]$.

Multiple features previously shown to correlate with bradykinesia were extracted from the four time-series (x, y, z, and vector magnitude) (see extensive overview including references in table S1). The features included characteristics from the temporal domain, such as extreme values, variances, jerkiness, number of peaks, and root mean squares, and the spectral domain, such as spectral power in specific frequency ranges, and dominant frequencies. The standard window length of analysis for each extracted feature was set as 60 seconds, meaning one value per feature was extracted per time series over every 60 seconds of data. To explore the influence of varying window lengths (3, 10, 30, 90, 120, 150, and 300 seconds), separate feature sets were extracted for each sub analysis. The resulting individual feature sets were balanced for medication-

status. Per participant, the surplus of minutes of available data in the longest recording (pre- or post-medication) were discarded at the end of the longest recording. This led to equal numbers of minutes of data, and thus features, from pre- and post-medication recordings per participant. Features were standardised individually, by extracting the mean of only the pre-medication features from a value, and dividing the result by the standard deviation of only the pre-medication recordings [67].

Descriptive statistics and analysis of variance

The demographic and disease characteristics of the included participants are described in Table 1. Unilateral scores are provided only for the side of which accelerometer data is included. To first test statistical distinguishability of the pre- and post-medication recordings on a group, before using the entire dataset as an input, four main accelerometer features were chosen a priori. These four features covered the most often used domains of motion metrics applied for naturalistic bradykinesia monitoring (maximum acceleration, coefficient of variation of acceleration over time, root mean square of acceleration over time, and the total spectral power below 4 Hz) [18,37], and were extracted from the vector magnitude time series. Individual averages of each of the four features over the entire dataset (~60 minutes per condition) were analysed for statistically significant differences between the medication states with a multivariate analysis of variance (M-ANOVA). Post-hoc repeated measures ANOVA were performed to explore which feature(s) contributed to the pre- versus post-medication difference. An alpha-level of 0.05 was implemented and multiple comparison correction was performed using the false discovery rate (FDR) method described by Benjamini and Hochberg [68].

Classification of medication states - Individually trained and group trained models

Supervised classification analyses were performed to test whether differentiation between short-term pre- and post-medication was feasible, based on 60-second accelerometer features (figure 1). First, this was tested using the four previously mentioned features extracted from the vector magnitude signal, afterwards the feature set was expanded to include all described features, as well as the x, y, z time series (table S1). Analyses were performed using a support vector machine (SV) and a random forest (RF) classifier. Classification models trained on individual data and models trained on group data were then compared (figure 1B). For individually trained models, 80% of a participant's total data was used as training data, and 20% as test data. Both were balanced for pre- and post-medication data (figure S1). Small blocks (2%) of training data which neighboured the test data were discarded (figure S1). This was done to decrease the temporal dependence between training and test data. To prevent bias caused by the selected block of test data, a 41-fold cross-validation was performed. Each fold used different blocks as test data (figure S2).

For group trained models, a leave one out cross-validation was performed. For every participant, a model was trained based on all data (balanced for medication status) from the remaining 19 participants and tested on all data (balanced for medication status) of the specific participant (figure 1B). To assess all models, the area under the receiver operator curve (AUC) and the classification accuracy were calculated as predictive metrics. For the individual models, individual classification outcomes were averaged over the 41 folds.

Permutation tests with 5000 repetitions were performed to test statistical significance for both individual and group trained models. The 95th percentile of permutation scores was taken as significance threshold (alpha = 0.05), and FDR multiple comparison corrections were performed [68].

Activity filtering

The reported analyses were repeated after removing data windows without movement activity. To identify data windows that do not contain any motion activity, different methodologies of activity filtering are described in PD monitoring literature [9,37,69]. We applied an activity filter which classified every 60 seconds window with a coefficient of variation of the vector magnitude less than 0.3 as 'no activity' and discarded them from analysis (figure S2). The choice of selected feature is based on previous work }[37], and the threshold is chosen pragmatically by group-level observations of video-annotated sections identified as non-active [36]. The activity-filtered data sets were individually balanced for medication-states. For example, if a participant's data set resulted in 50 'active' minutes pre-medication, and only 45 'active' minutes post-medication, the surplus of features from 5 'active' minutes pre-medication were discarded at the end of the data set. On

average, 44.5 minutes (+/- 13.9 minutes) of features were included after applying the activity filter and balancing the individual data to include equal individual features per medication state (Table 1).

The influence of training data size, and feature window lengths
To test the impact of the size of the training set in the group models, the training phases were repeated with varying numbers of participants included in the training data (figure 4). As in the original group model analysis, the test data consisted of all data from one participant. The number of training data participants varied between 1 and 19. To prevent selection bias in the selection of the training participants, the analyses were repeated five times per number of included training participants, with different random selections of training participants. Individual models were excluded from this analysis.
To analyse the influence of feature window lengths, we repeated the group model analysis with features extracted from data windows of 3, 10, 30, 90, 120, 150, and 300 seconds duration (figure 4). For every analysis, one participant was selected as test participant, and the other 19 were training participants. This was repeated for all subjects and the average results over 20 test participants were reported. This was performed at the group level modelling only, as individual models were limited by total available data size.

Comparing two models' predictive performance
Equality plots were drawn to compare the predictive performance between two models (figure S3). For example, model A led to better results in 14 out of 20 participants than model B. Permutation tests plotted 20 random dots on an equality plot and tested whether the permuted distribution generated 14 or more dots (out of 20) above the equality line. This was repeated 5000 times, and the probability that the distribution '14 out of 20' was the result of chance was determined.

Predictive performance and clinical assessed symptom fluctuations
The influence of clinical bradykinesia and tremor fluctuations on predictive performance was tested at a group level by Spearman R correlations between the fluctuation in individual bradykinesia and tremor sub scores, and the predictive performance (table S3). Individual participants were visualized according to descending tremor fluctuation ratings to enable visual comparison of predictive performance with and without co-occurring tremor fluctuation (figure S4). The tremor scores consisted of the MDS-UPDRS III items representing unilateral upper extremity tremor (items 15b, 16b, and 17b for the left side, and items 15a, 16a, and 17a for the right side).

Software
Raw acceleration time series were down sampled and filtered (for gravity effects) in Matlab. All further pre-processing, feature extraction, and analysis was performed in Jupyter Notebook (Python 3.7). The code used to extract features and analyse data is available at www.github.com/jgvhabets/brady_reallife/ [70].

Financial disclosures
JH received funding from the Dutch health care research organization ZonMW (Translational Research 2017 - 2024 grant nr. 446001063). JH, PT and YT received funding from the Stichting Weijerhorst. CH received a VENI-funding from the Dutch Research Council (WMO).

Supplementary Material

Feature name	Source/ literature
Temporal domain	
Maximum acc	71
Interquartile range of acc	71
90th percentile of acc	72
median of acc	69,73
mean of acc	69,73
standard deviation of acc	33
variance of acc	33
coefficient of variation	33
acceleration range	37
low acc peaks (n)	74
high acc peaks (n)	74
time spent above 1g acc (%)	69,75
acc entropy	34,37
jerkiness ratio/ smoothness	34,37
root mean square (RMS)	37
Ratio of x/y/z-RMS compared to vector magnitude-RMS	76
Axial cross-correlation (X-Y; (X-Z,; Y-Z)	34,37
Spectral domain	
spectral power < 3.5 Hz	36,71
spectral power 0.7 < 1.4 Hz	36,71
spectral power 1.4 < 2.8 Hz	36,71
spectral power 2.8 < 3.5 Hz	36,71
spectral flatness	37
spectral entropy	37
spectral variance	74,77
spectral smoothness	74
spectral low/high peaks	74
Dominant frequency magnitude	34,37
Dominant frequency ratio	37
Dominant frequency flatness	37
Dominant frequency entropy	37

Table S1: Extracted features over x-, y-, z-, axes, and signal vector magnitude time-series.

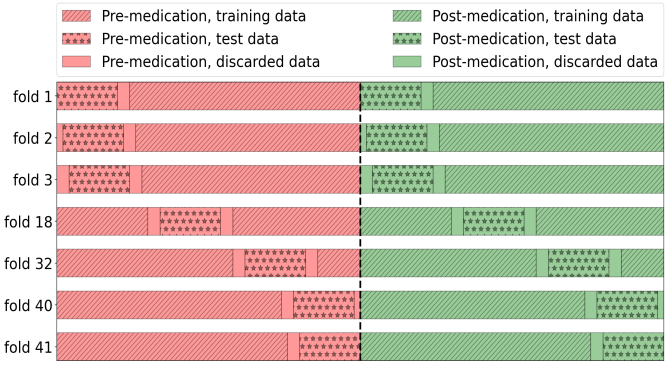


Figure S1: Schematic visualization of data splitting method for individual models. Pre- and post-medication features were balanced in number. ⅛ of pre-medication data, and ⅛ of post-medication data, 20% of total data, were selected as 'hold-out' validation data for the test phase. From the remaining 80%, the 2 adjacent percentages of data were not included in the training data, to decrease the temporal dependence of the training and test data. Folds 1, 2, 3, 18, 32, 40, and 41 are shown as an example.

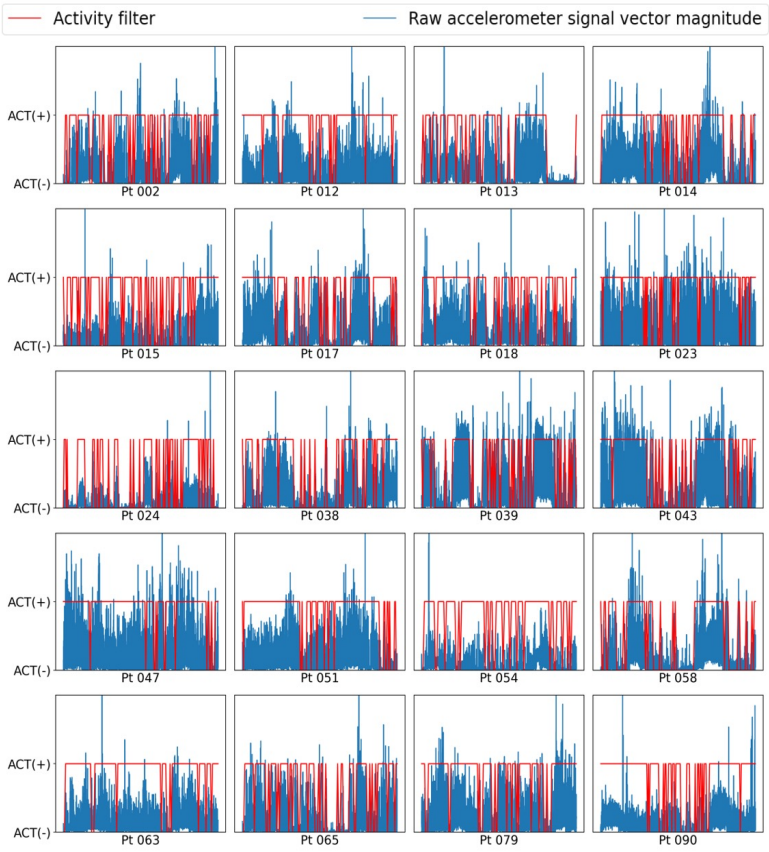


Figure S2: Visualization of activity filter performance versus the parallel raw signal vector magnitude. ACT(+): activity filter indicated activity, ACT(+): activity filter indicated no activity. Pt: participant.

			All features	All features		4 features	4 features
CLASSIFIER		n=20, mean (sd)	All minutes	Activity filtered		All minutes	Activity filtered
SUPPORT VECTOR	INDIVIDUAL MODEL	auc	0.682 (0.15)	0.696 (0.18)		0.499 (0.12)	0.533 (0.16)
		auroc, n sign	16	13		2	4
		accuracy	0.632 (0.12)	0.651 (0.14)		0.490 (0.11)	0.509 (0.14)
		accuracy, n sign	15	14		5	10
	GROUP MODEL	auroc	0.669 (0.10)	0.703 (0.10)		0.590 (0.11)	0.633 (0.13)
		auroc, n sign	16	17		10	11
		accuracy	0.624 (0.09)	0.640 (0.08)		0.560 (0.11)	0.597 (0.08)
		accuracy, n sign	11	12		9	12
RANDOM FOREST	INDIVIDUAL MODEL	auroc	0.649 (0.13)	0.656 (0.17)		0.586 (0.12)	0.619 (0.14)
		auroc, n sign	15	10		6	11
		accuracy	0.611 (0.10)	0.611 (0.13)		0.558 (0.10)	0.588 (0.11)
		accuracy, n sign	13	10		8	7
	GROUP MODEL	auroc	0.661 (0.10)	0.698 (0.11)		0.593 (0.11)	0.636 (0.14)
		auroc, n sign	12	16		10	10
		accuracy	0.598 (0.08)	0.626 (0.08)		0.564 (0.08)	0.587 (0.08)
		accuracy, n sign	12	12		6	7

Table S2: Predictive performance of different models. All scores are group averages over twenty participants per model. Sd: standard deviation.

Figure S3: Comparison of different classification model approaches. Equality plots comparing two models regarding individual area under the receiver operator characteristic (AUC) and accuracy. Each dotted line visualizes the line x = y, and represents equality of the two displayed models. SV: support vector classifier, RF: random forest classifier.
The p-values throughout figure S3 indicate whether the ratio of patients that scored higher on model X versus model Y is statistically significant. We performed a 5000 permutation test where 20 dots (random x-value, random y-value) were randomly plotted in the equality plot. The p-values represent the chance that the distribution is better than random chance level.

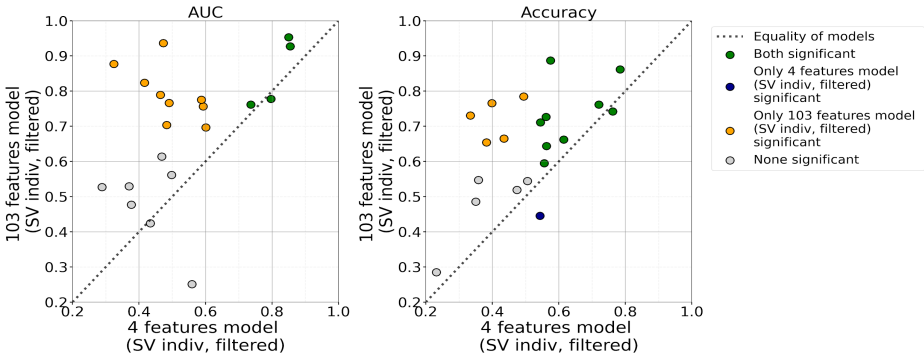


Figure S3A: Individual models: 4 versus 103 features. AUC: 17 out of 20 higher, $p < 0.000$, accuracy: 18 out of 20 higher: $p < 0.000$

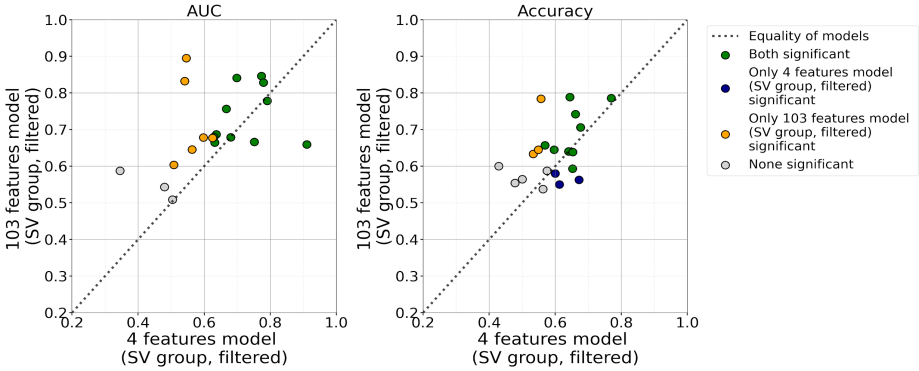


Figure S3B: Group models: 4 versus 103 features. AUC: 16 out of 20 higher, $p = 0.002$, accuracy: 14 out of 20 higher: $p = 0.023$

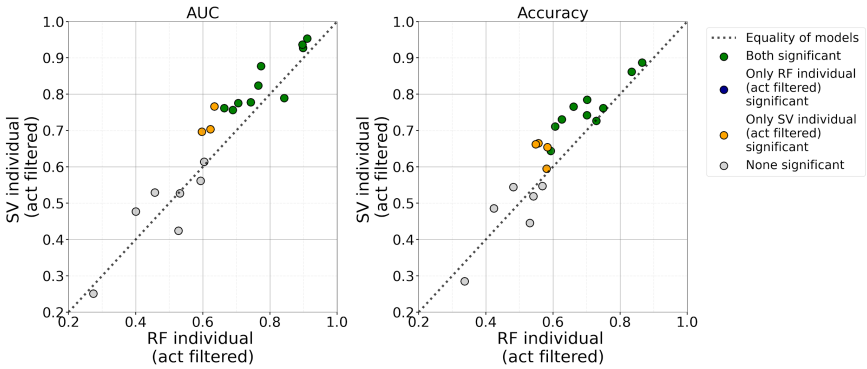


Figure S3C: Individual models: SV versus RF classifiers. AUC: 15 out of 20 higher, p = 0.009, accuracy: 15 out of 20 higher: p = 0.009

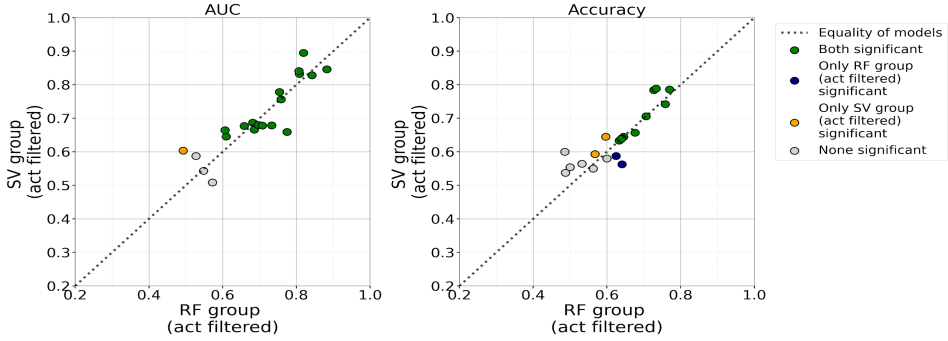


Figure S3D: Group models: SV versus RF classifiers. AUC: 10 out of 20 higher, p = 0.406, accuracy: 10 out of 20 higher: p = 0.406

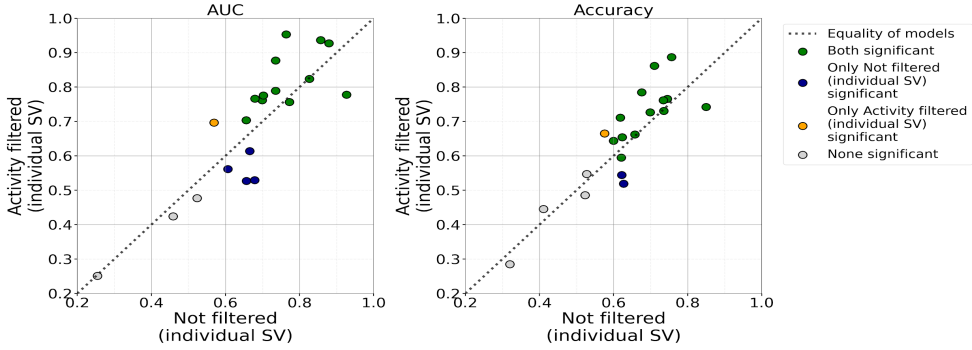


Figure S3E: Individual SV models: with activity filtering versus without activity filtering. AUC in 10 out of 20 higher, p = 0.406, accuracy in 13 out of 20 higher, p = 0.056. Note that both the AUC scores and the classification accuracies of the activity filtered models are mainly higher than those of the not filtered models, when all 'none significant' candidates are disregarded. We conclude that although there is no statistically significant superiority of the activity filtered models, there is a trend that activity filtered models lead to higher predictive performance.

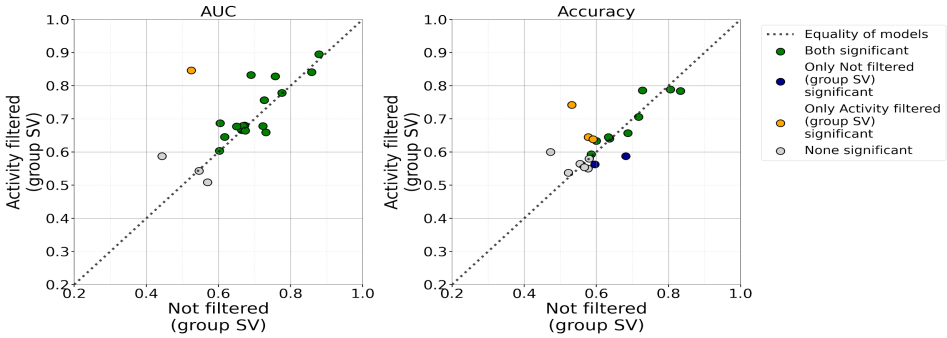


Figure S3F: Group SV models: with activity filtering versus without activity filtering. AUC in 13 out of 20 higher, p = 0.056, accuracy in 11 out of 20 higher, p = 0.250.

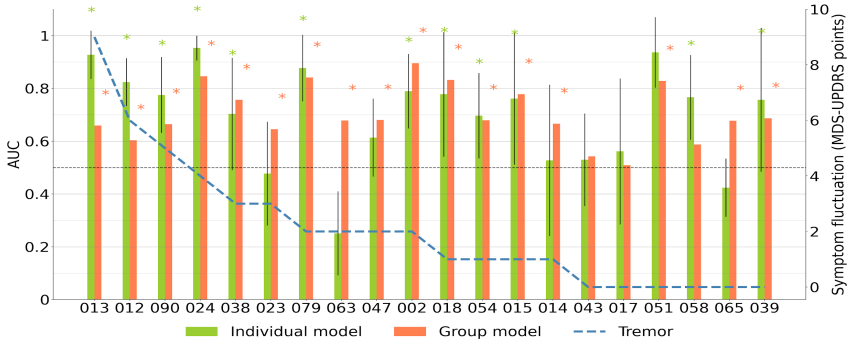


Figure S4: Good classification performance in patients with and without tremor

Colored bars visualize the individual AUC scores from the best individual and the best group model (both support vector classifier, and activity filtered), and correspond to the left y-axis. Individual tremor fluctuations between pre- and post-medication correspond to the right y-axis. Tremor scores represent the described MDS-UPDRS III items for unilateral upper extremity tremor (see Methods). On the x-axis individual participants are sorted on tremor fluctuation, in descending order.

Colored asterisks indicate statistical significance of the AUC score compared to chance level (alpha = 0.05, FDR corrected). The black dotted line indicates chance-level for the AUC scores. AUC: area under the receiver operator characteristic; FDR: false discovery rate, MDS-UPDRS: Movement Disorders Society - Unified Parkinson Disease Rating Scale.

	Individual models, AUC	Group models, AUC	Individual models, accuracy	Group models, accuracy
Bradykinesia (r (p))	0.24 (p = 0.305)	0.01 (p = 0.962)	0.24 (p = 0.305)	0.18 (p = 0.452)
Tremor (r (p))	0.34 (p = 0.140)	0.11 (p = 0.642)	0.21 (p = 0.380)	-0.06 (p = .807)

Table S3: Spearman R correlations between symptom fluctuation and predictive performance at an individual level. Spearman r correlations are calculated between the MDS-UPDRS tremor and bradykinesia fluctuation as described in the Methods, and the predictive performance per participant. Support vector models including activity filtering were compared for individual and group model comparisons.

References

1. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet Lond. Engl.* (2021) doi:10.1016/S0140-6736(21)00218-X.
2. Jankovic, J. & Tan, E. K. Parkinson's disease: etiopathogenesis and treatment. *J Neurol Neurosurg Psychiatry* **91**, 795–808 (2020).
3. Kuhlman, G. D., Flanigan, J. L., Sperling, S. A. & Barrett, M. J. Predictors of health-related quality of life in Parkinson's disease. *Parkinsonism Relat. Disord.* **65**, 86–90 (2019).
4. de Bie, R. M. A., Clarke, C. E., Espay, A. J., Fox, S. H. & Lang, A. E. Initiation of pharmacological therapy in Parkinson's disease: when, why, and how. *Lancet Neurol.* **19**, 452–461 (2020).
5. Fasano, A. *et al.* Characterizing advanced Parkinson's disease: OBSERVE-PD observational study results of 2615 patients. *BMC Neurol* **19**, 50 (2019).
6. Kim, H.-J. *et al.* Motor Complications in Parkinson's Disease: 13-Year Follow-up of the CamPaIGN Cohort. *Mov. Disord.* **35**, 185–190 (2020).
7. Hechtner, M. C. *et al.* Quality of life in Parkinson's disease patients with motor fluctuations and dyskinesias in five European countries. *Parkinsonism Relat. Disord.* **20**, 969–974 (2014).
8. Fox, S. H. *et al.* International Parkinson and movement disorder society evidence-based medicine review: Update on treatments for the motor symptoms of Parkinson's disease. *Mov Disord* **33**, 1248–1266 (2018).
9. van Hilten, J. J., Middelkoop, H. A., Kerkhof, G. A. & Roos, R. A. A new approach in the assessment of motor activity in Parkinson's disease. *J Neurol Neurosurg Psychiatry* **54**, 976–9 (1991).
10. Espay, A. J. *et al.* A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Mov Disord* **34**, 657–663 (2019).
11. Warmerdam, E. *et al.* Long-term unsupervised mobility assessment in movement disorders. *Lancet Neurol.* **19**, 462–470 (2020).
12. Fasano, A. & Mancini, M. Wearable-based mobility monitoring: the long road ahead. *Lancet Neurol.* **19**, 378–379 (2020).
13. Odin, P. *et al.* Viewpoint and practical recommendations from a movement disorder specialist panel on objective measurement in the clinical management of Parkinson's disease. *NPJ Park. Dis* **4**, 14 (2018).
14. Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord* **23**, 2129–70 (2008).
15. Hagell, P. & Nygren, C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. *J Neurol Neurosurg Psychiatry* **78**, 1191–8 (2007).
16. Hauser, R. A., McDermott, M. P. & Messing, S. Factors associated with the development of motor fluctuations and dyskinesias in Parkinson disease. *Arch. Neurol.* **63**, 1756–1760 (2006).
17. Papapetropoulos, S. S. Patient diaries as a clinical endpoint in Parkinson's disease clinical trials. *CNS Neurosci Ther* **18**, 380–7 (2012).
18. Thorp, J. E., Adamczyk, P. G., Ploeg, H. L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. *Front Neurol* **9**, 1036 (2018).
19. Horne, M. K., McGregor, S. & Bergquist, F. An objective fluctuation score for Parkinson's disease. *PLoS One* **10**, e0124522 (2015).
20. Sama, A. *et al.* Estimating bradykinesia severity in Parkinson's disease by analysing gait through a waist-worn sensor. *Comput Biol Med* **84**, 114–123 (2017).
21. Powers, R. *et al.* Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease. *Sci. Transl. Med.* **13**, eabd7865 (2021).
22. Sica, M. *et al.* Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review. *PLOS ONE* **16**, e0246528 (2021).
23. Pahwa, R. *et al.* Role of the Personal KinetiGraph in the routine clinical assessment of Parkinson's disease: recommendations from an expert panel. *Expert Rev Neurother* **18**, 669–680 (2018).
24. Santiago, A. *et al.* Qualitative Evaluation of the Personal KinetiGraph TM Movement Recording System in a Parkinson's Clinic. *J. Park. Dis.* **9**, 207–219 (2019).
25. Joshi, R. *et al.* PKG Movement Recording System Use Shows Promise in Routine Clinical Care of Patients With Parkinson's Disease. *Front Neurol* **10**, 1027 (2019).
26. Farzanehfar, P., Woodrow, H. & Horne, M. Assessment of Wearing Off in Parkinson's disease using objective measurement. *J Neurol* (2020) doi:10.1007/s00415-020-10222-w.

27. Ossig, C. *et al.* Correlation of Quantitative Motor State Assessment Using a Kinetograph and Patient Diaries in Advanced PD: Data from an Observational Study. *PLoS One* **11**, e0161559 (2016).
28. Rodriguez-Molinero, A. *et al.* Validation of a portable device for mapping motor and gait disturbances in Parkinson's disease. *JMIR Mhealth Uhealth* **3**, e9 (2015).
29. Rodriguez-Molinero, A. Monitoring of Mobility of Parkinson's Patients for Therapeutic Purposes - Clinical Trial (MoMoPa-EC). NCT04176302 (2019).
30. Great Lake Technologies. Kinesia 360 Parkinson's Monitoring Study. (2018).
31. van Halteren, A. D. *et al.* Personalized Care Management for Persons with Parkinson's Disease. *J. Park. Dis.* **10**, S11–S20 (2020).
32. Clarke, C. E. *et al. UK Parkinson's Disease Society Brain Bank Diagnostic Criteria*. *Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate Parkinson's disease: a large pragmatic randomised controlled trial (PD REHAB)* (NIHR Journals Library, 2016).
33. Shawen, N. *et al.* Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. *J. NeuroEngineering Rehabil.* **17**, 52 (2020).
34. Lonini, L. *et al.* Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *Npj Digit. Med.* **1**, 64 (2018).
35. Fisher, J. M. *et al.* Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. *Parkinsonism Relat. Disord.* **33**, 44–50 (2016).
36. Evers, L. J. *et al.* Real-Life Gait Performance as a Digital Biomarker for Motor Fluctuations: The Parkinson@Home Validation Study. *J Med Internet Res* **22**, e19068 (2020).
37. Mahadevan, N. *et al.* Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *NPJ Digit Med* **3**, 5 (2020).
38. Khodakarami, H. *et al.* Prediction of the Levodopa Challenge Test in Parkinson's Disease Using Data from a Wrist-Worn Sensor. *Sensors* **19**, 5153 (2019).
39. Velisar, A. *et al.* Dual threshold neural closed loop deep brain stimulation in Parkinson disease patients. *Brain Stimulat.* **12**, 868–876 (2019).
40. Castaño-Candamil, S. *et al.* A Pilot Study on Data-Driven Adaptive Deep Brain Stimulation in Chronically Implanted Essential Tremor Patients. *Front. Hum. Neurosci.* **14**, (2020).
41. Gilron, R. *et al.* Long-term wireless streaming of neural recordings for circuit discovery and adaptive stimulation in individuals with Parkinson's disease. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00897-5.
42. Habets, J. G. V. *et al.* An update on adaptive deep brain stimulation in Parkinson's disease. *Mov Disord* **33**, 1834–1843 (2018).
43. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 195 (2019).
44. Heijmans, M., Habets, J., Kuijf, M., Kubben, P. & Herff, C. Evaluation of Parkinson's Disease at Home: Predicting Tremor from Wearable Sensors. in 584–587 (2019). doi:10.1109/EMBC.2019.8857717.
45. Heijmans, M. *et al.* Monitoring Parkinson's disease symptoms during daily life: a feasibility study. *NPJ Park. Dis* **5**, 21 (2019).
46. Vizcarra, J. A. *et al.* The Parkinson's disease e-diary: Developing a clinical and research tool for the digital age. *Mov Disord* (2019) doi:10.1002/mds.27673.
47. Habets, J. *et al.* Mobile Health Daily Life Monitoring for Parkinson Disease: Development and Validation of Ecological Momentary Assessments. *JMIR Mhealth Uhealth* **8**, e15628 (2020).
48. Habets, J. G. V. *et al.* A Long-Term, Real-Life Parkinson Monitoring Database Combining Unscripted Objective and Subjective Recordings. *Data* **6**, 22 (2021).
49. Versi, E. 'Gold standard' is an appropriate term. *Bmj* **305**, 187 (1992).
50. Galperin, I. *et al.* Associations between daily-living physical activity and laboratory-based assessments of motor severity in patients with falls and Parkinson's disease. *Park. Relat Disord* **62**, 85–90 (2019).
51. Pencina, M. J., Goldstein, B. A. & D'Agostino, R. B. Prediction Models - Development, Evaluation, and Clinical Application. *N Engl J Med* **382**, 1583–1586 (2020).
52. Bloem, B. R. *et al.* The Personalized Parkinson Project: examining disease progression through broad biomarkers in early Parkinson's disease. *BMC Neurol* **19**, 160 (2019).
53. J. Watts, A. Khojandi, R. Vasudevan, & R. Ramdhani. Optimizing Individualized Treatment Planning for Parkinson's Disease Using Deep Reinforcement Learning. in *2020 42nd Annual International Conference of the*

IEEE Engineering in Medicine & Biology Society (EMBC) 5406–5409 (2020). doi:10.1109/EMBC44109.2020.9175311.
54. Matias, R., Paixão, V., Bouça, R. & Ferreira, J. J. A Perspective on Wearable Sensor Measurements and Data Science for Parkinson's Disease. *Front. Neurol.* **8**, 677 (2017).
55. Little, S., Pogosyan, A., Kuhn, A. A. & Brown, P. beta band stability over time correlates with Parkinsonian rigidity and bradykinesia. *Exp Neurol* **236**, 383–8 (2012).
56. Rochester, L. *et al.* A Roadmap to Inform Development, Validation and Approval of Digital Mobility Outcomes: The Mobilise-D Approach. *Digit. Biomark.* **4**, 13–27 (2020).
57. Kluge, F. *et al.* Consensus based framework for digital mobility monitoring. *medRxiv* 2020.12.18.20248404 (2020) doi:10.1101/2020.12.18.20248404.
58. MJFF, S. BEAT-PD DREAM Challenge (by Sage Bionetworks; Michael J. Fox Foundation). (2020).
59. Pulliam, C. L. *et al.* Continuous Assessment of Levodopa Response in Parkinson's Disease Using Wearable Motion Sensors. *IEEE Trans Biomed Eng* **65**, 159–164 (2018).
60. Daneault, J. *et al.* Estimating Bradykinesia in Parkinson's Disease with a Minimum Number of Wearable Sensors. in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* 264–265 (2017). doi:10.1109/CHASE.2017.94.
61. Daneault, J.-F. *et al.* Accelerometer data collected with a minimum set of wearable sensors from subjects with Parkinson's disease. *Sci. Data* **8**, 48 (2021).
62. Hua, R. & Wang, Y. Monitoring Insole (MONI): A Low Power Solution Toward Daily Gait Monitoring and Analysis. *IEEE Sens. J.* **19**, 6410–6420 (2019).
63. Lima, A. L. S. de *et al.* Large-Scale Wearable Sensor Deployment in Parkinson's Patients: The Parkinson@Home Study Protocol. *JMIR Res. Protoc.* **5**, e5990 (2016).
64. Lima, A. L. S. de *et al.* Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease. *PLOS ONE* **12**, e0189161 (2017).
65. Kim, S. J., Koh, K. ,. Boyd, S. ,. Gorinevsky, D. l(1) Trend Filtering. *SIAM Rev* **May 51**, 339–360 (2009).
66. van Brummelen, E. M. J. *et al.* Quantification of tremor using consumer product accelerometry is feasible in patients with essential tremor and Parkinson's disease: a comparative study. *J Clin Mov Disord* **7**, 4 (2020).
67. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
68. Korthauer, K. *et al.* A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 118 (2019).
69. Keijsers, N. L., Horstink, M. W. & Gielen, S. C. Ambulatory motor assessment in Parkinson's disease. *Mov Disord* **21**, 34–44 (2006).
70. jgvhabets/brady_reallife: First release for short-term, individual and group modelling analyses | Zenodo. https://zenodo.org/record/4734199#.YJAOZRQza3J.
71. Griffiths, R. I. *et al.* Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. *J Park. Dis* **2**, 47–55 (2012).
72. Rispens, S. M. *et al.* Identification of fall risk predictors in daily life measurements: gait characteristics' reliability and association with self-reported fall history. *Neurorehabil. Neural Repair* **29**, 54–61 (2015).
73. Hoff, J. I., van der Meer, V. & van Hilten, J. J. Accuracy of objective ambulatory accelerometry in detecting motor complications in patients with Parkinson disease. *Clin Neuropharmacol* **27**, 53–7 (2004).
74. Balasubramanian, S., Melendez-Calderon, A. & Burdet, E. A robust and sensitive metric for quantifying movement smoothness. *IEEE Trans. Biomed. Eng.* **59**, 2126–2136 (2012).
75. Salarian, A. *et al.* Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system. *IEEE Trans Biomed Eng* **54**, 313–22 (2007).
76. Sekine, M. *et al.* A gait abnormality measure based on root mean square of trunk acceleration. *J. NeuroEngineering Rehabil.* **10**, 118 (2013).
77. Beck, Y. *et al.* SPARC: a new approach to quantifying gait smoothness in patients with Parkinson's disease. *J. NeuroEngineering Rehabil.* **15**, 49 (2018).

Chapter 10

General discussion

The aim of this thesis is to improve future deep brain stimulation (DBS) care for Parkinson patients via preoperative outcome prediction and real-life (naturalistic) symptom monitoring. In part A, we demonstrate a proof-of-principle of individual preoperative DBS outcome prediction. The development and external validation of the prediction model (chapters 2 and 3) can be part of the foundation of future clinical decision support systems (CDSS) for DBS care. In part B, we first explore how subjective experience sampling methods (ESM) can attribute to naturalistic Parkinson monitoring (chapters 5 to 8). Then we discuss the feasibility and possible optimization methodologies of naturalistic motor monitoring over short time windows based on wearable motion sensors. Finally, we will discuss the potential role of short-term naturalistic motor monitoring in adaptive DBS evaluation or programming.
The clinical data science methodologies applied throughout this thesis can positively influence clinical DBS practice. To ensure this positive clinical impact [1], clinical expertise needs to guide the computational development and validation, and computational expertise needs to guide the clinical interpretation and implementation. Fortunately, specific guidelines are becoming available to help this multidisciplinary effort [2–4]. We see it as a responsibility of the present generation clinical neuroscientists and scientifically engaged neurologists and neurosurgeons to bridge the gap between data scientists and clinicians and ensure the modern technological possibilities are translated in value for the patient.

10.1 Creating individual preoperative STN DBS outcome prediction with DBS-PREDICT

Development and external validation of DBS-PREDICT and remaining challenges
Individual subthalamic nucleus (STN) DBS outcome prediction and preoperative identification of the suboptimal responding minority [5] is a topic of interest for nearly two decades [6]. Machine learning driven prediction models try to find patterns to identify that patient population, which are not findable without these computational analyses. The proof-of-concept that these discriminating patterns are present in preoperative clinical variables, is described in chapter 2, and by Frizon et al [7]. The created models generate individual outcome predictions based on individual clinical variables, in contrast to the existing literature which unravelled correlations between preoperative variables and outcome on a group level [8–10]. In chapter 3 we demonstrated that the learned patterns of the machine learning model from chapter 2 (DBS-PREDICT) were able to identify the suboptimal responders in an external population, independent from the model development. Such an external validation is regarded as the gold standard for model evaluation and is necessary steps after model development [11]. So far, no comparable external validation study of a preregistered outcome prediction model has been published in DBS care. Therefore, this work can be regarded as an important step towards an impactful clinical decision support system (CDSS) for individual outcome prediction in DBS care. Here, we will elaborate on essential remaining steps which we believe have to be taken before impactful clinical implementation can be realized.

Finding an automated outcome definition
One of the biggest challenges inherent to machine learning based DBS outcome prediction is finding a widely accepted, automated outcome definition. A quick look at some established effectivity studies for DBS in PD learns us that several measures are used (in different therapeutic states) such as (MDS) UPDRS total scores, (MDS) UPDRS sub scores, MDS UPDRS single item scores, QoL-scales, and levodopa equivalent daily dosage changes [5,12–15]. Finding an outcome

definition suitable for automated classification (or regression) therefore requires arbitrary choices on which outcome measures to include, which cut off thresholds to use, and how to combine these outcome measures. We argue that an outcome definition suitable for individual, automated DBS outcome prediction does not need to correlate perfectly with existing, single, outcome measures. It is of more importance that it correlates reasonably well, and that its clinical implementation will add clinical value to the patient's care.
In chapter 2 and 3, we defined binary outcome classes by applying literature-based cut off values for change in motor symptoms, adverse events, and daily life function (in optimal therapeutic condition). Although we aimed for a holistic definition, our approach can be disputed. Motor symptoms in medication-off condition also could have been considered, or specific symptom related (MDS) UPDRS single items, or QoL-scales. Moreover, we chose for a negative outcome (weak responder), where a positive outcome also would have been possible, albeit for a different clinical utilization (see next paragraph). Concluding this, we argue that our outcome classification is sufficient to demonstrate proof-of-concept and to pave the way for the remaining scientific steps, especially given the data limitations inherent to retrospective (multicenter) cohorts.

Creating a clinical utilization with optimal impact on STN-DBS care
Generally, the trained machine learning on which a CDSS is build has to answer a specific clinical question in a specific population, but should be generalizable enough to work in external, comparable, populations (and violation of the latter is called 'overfitting'). Following these principles, the exact clinical utilization of a CDSS has to be clear on the moment of development study design. Therefore, our work in chapters 2 and 3 is a valuable step towards clinically impactful CDSS for DBS care, both despite of and thanks to its limitations.
We developed and validated DBS-PREDICT among a population who underwent STN DBS. Hence, the model cannot be applied on a population who is referred for multidisciplinary DBS consultation to detect potential strong or weak responder. It would require a new development study including all PD patients referred for DBS consultation, to create a CDSS machine learning model for this clinical question. Also, the clinical question can differ. Instead of detecting weak responders, as DBS-PREDICT does, a CDSS could detect strong responders. Besides training the model with a different outcome classification, this would have influences for the statistical interpretation of true and false positive and negative values.
The latter interpretation can even differ within the same model development study, depending on the desired utilization in clinical practice. As explained in chapter 2 and 3 (see for example table 1 and figure 4 in chapter 3), the exact predictive performance of a specific scenarios (different probability thresholds) can have different clinical and socioeconomic effects [16]. For example, chapter 3 describes strong negative ('negative prediction' is coded as 'strong responder') predictive values. These high predictive values of 'strong response' predictions (0.79 to 0.84) could serve the following scenario. When a clinician is in doubt during clinical counselling, and DBS-PREDICT generates a strong response, this can reassure the clinician to include the patient for STN DBS. In case DBS-PREDICT generates a 'weak response' prediction, this will only be correct in 49% to 62% of the cases. The clinician will be extra warned to trust any gut feeling, and to explicitly discuss this with the patient and the multidisciplinary team.
In any case, individual patient-preferences or -circumstances can influence the shared decision-making process. The results and interpretations of CDSS outcomes should be translated meticulously to actually support the clinician and the patient in this crucial decision-making.

Therefore, a model as DBS-PREDICT is intended to advice the clinician, rather than entirely overtake the decision-making process from the clinician.

Demonstrating predictive validity in a prospective, observational, preregistered study
To better understand the actual clinical impact of a CDSS on decision making in DBS care, we propose an observational prospective study without interference of current regular care. To ensure generalizability a preregistered multicenter study is desirable. This could be realized by providing an independent clinical expert without a role in the decision making with all patient information and the additional CDSS result and compare the therapeutic advice. This would provide information about the interaction between data driven CDSS, clinician and potentially even patient. This understanding of human (clinician) behaviour affected by CDSS will likely play an important in future CDSS development [4]. Based on the prospective predictive performance of the model, and the results on clinical decision-making interaction, a clinical utilization with optimal clinical impact could be designed.
Moreover, prospective data collection would enable the inclusion of lacking preoperative clinical predictors and postoperative outcome values. Also, our binary outcome classification can be validated against QoL-scores. The inclusion of additional preoperative clinical predictors can be considered, and models including and excluding these additional predictors can be compared. Additional data types can include kinematic data [17], structural imaging [18], connectivity imaging [19], preoperative neurophysiology [20], and genetic profiling [21]. While making these decisions, the burden for the patient and the usability and generalizability of the CDSS has to be kept in mind.

10.2 Lessons learned regarding individual DBS outcome prediction

Considering the anticipated increasing role of data driven CDSS in neurological and neurosurgical practice [22], a thorough understanding of this new multidisciplinary field is relevant. Collaboration, communication and mutual understanding between clinicians and data scientists is essential to ensure clinical impact [16,23], and specific guidelines are available to support this effort [1,3,24].
Especially given the complicated multifactorial, and partly unravelled, mechanisms of action in DBS [25], a structured scientific approach is necessary to realise CDSS-supported individual DBS outcome prediction. So far, the literature on individual outcome prediction in DBS care lacked this structured methodology, and consequently some important considerations were missing. We argue that every scientific work discussing individual DBS outcome prediction should consider the broader view on CDSS development and implementation, as discussed in the previous paragraphs. We believe our results and discussion can catalyse the individual outcome prediction development for DBS, and we are looking forward to seeing first results of prospective CDSS studies in DBS practice. We encourage all researchers in this field to publish their study protocols. As a transparent open-science attitude will improve scientific progress in this pioneering field [26].

10.3 Naturalistic real-life Parkinson monitoring

Experience sampling method (ESM) as part of naturalistic Parkinson monitoring
This thesis explored the potential of ESM to improve naturalistic Parkinson monitoring. In chapter 5, we demonstrated the feasibility of Parkinson monitoring combining motion sensors and smartphone-based ESM over two weeks. ESM completion rates and wearing time of the motion sensors were high, and participants reported to use the devices regularly if it would benefit their therapy management. Then, chapter 6 showed that the applied PD-specific ESM questionnaire resulted in valuable data regarding motor symptomatology and function. The negative correlations between symptom severity and motor functionality and positive correlation between symptom severity and off-medication support the questionnaire's structural and internal validity. However, replication and validation of these results in a larger PD population with known fluctuations, including gold standard assessments, is indicated. Since there is no gold standard with a similar assessment frequency, pragmatic and creative study designs are required for this. ESM answers could be compared over a longer term with MDS-UPDRS scores from the same period of time for momentary validity. A comparison with Hauser diaries and medication intake moments could assess the validity of fluctuation detection over time. After such validation, ESM for PD could provide long-term monitoring for general neurological practice and pharmacological clinical trials [27].
Similar validation studies can be done including validated psychiatric assessments to explore ESM's clear potential for non-motor symptom monitoring. This might even contribute to psychiatric adverse effects monitoring after DBS [1]. In general, real-life non-motor monitoring is lagging behind, parallel to the notorious neglect of non-motor symptoms in clinical PD practice [28].

In order to catalyse PD research using ESM recordings with or without motion sensing recordings, we made the data discussed in chapters 5, 6, and 8, publicly available [29]. Chapter 7 describes practical and conceptual details of the combined data set and provides an example how to combine and analyse the data.

ESM to improve naturalistic motor symptom detection in PD
Our findings from chapter 5, 6, and 8 suggest ESM can help to overcome the lack of a repetitive, ecological, gold standard assessment for motion sensor development and validation. Despite the many recent advances in naturalistic motor monitoring [30,31], most validation processes are criticised [32]. Current best practices are the comparison of months of monitoring data with a clinician's overall impression based on patient files [33], and the prospective comparison of patient outcomes after neurological out-patient clinic decision-making with and without additional multi-day monitoring [34]. Although these are valuable, creative examples how to overcome this challenge, the disputed within-day validity and short-term-validity remain unanswered [35].
Therefore, we hypothesize that ESM has the potential to contribute to development and validation of naturalistic PD motor monitoring by yielding real-life 'ground truth labels' multiple times daily over substantial periods of time. Chapter 8 provides a n=1 proof-of-concept of this hypothesis. Using ESM data as ground truth labels, we successfully developed an accelerometer-based detection model for naturalistic short-term tremor fluctuations. This method needs replication in larger populations and should be expanded with bradykinesia and dyskinesia

detection. Besides for model development, ESM data could also be used as ground truth labels for validation of already developed motor symptom detection models.

Feasibility of short-term, naturalistic PD motor monitoring

To investigate the potential of naturalistic motor monitoring for real-time and closed-loop feedback applications, we explored short-term validity of PD monitoring in chapter 9. We show that medication-induced naturalistic motor fluctuations can be differentiated based on accelerometer-derived bradykinesia features, over 60-second windows. The hypothesized additional value of personalized, individually trained models over group models was not found. Additional analyses with the group models showed that feature windows up to 300 seconds and larger data sets increased the predictive performance. The latter findings are in line with previous literature [36] and encourage replication of the individual model analyses with larger individual data sets.

In general, the predictive performance over 60-second windows support our hypothesis that naturalistic motor symptom detection is possible on a high temporal resolution. As discussed, despite high temporal resolutions of several commercial devices' symptom scores (seconds to minutes), aggregation of scores over longer periods of time is needed for good performance [33,35]. An important limitation is the individual variability of predictive performance, most prominently in the individualized models. Currently, we hypothesize that this variability is caused by the relatively small individual data sets and the inherent variability of the unscripted activities of daily life (ADL) recorded. The test data selected to compute the models' performance originated from blocks of 10 minutes. High variability in performed activities could lead to mismatching training and test data, resulting in poorer results [36,37]. However, this requires further analysis in larger or more structured individual data sets. As an intermediate step between scripted experiments and naturalistic experiments, video-observed ADL studies in movement labs can be considered.

Possibilities and limitations for short-term naturalistic PD motor monitoring

Regardless of the temporal resolution, future development and validation of PD monitoring devices should pay closer attention to the heterogenous PD symptomatology. Clear definitions of clinical utilization are needed, and in case they cover multiple symptoms, strategies to aggregate symptom scores have to be studied. In doing so, the intended interaction between the clinician and the provided scores has to be discussed [4].

Methodological boundaries of naturalistic short-term motor monitoring have to explored. Clinical utilizations have to be tailored to the minimal time window required for valid symptom detection. For this, individual models, and the in chapter 9 suggested optimizations, should be explored in larger, naturalistic, individual data sets.

So far, the literature nearly exclusively describes PD monitoring based on supervised machine learning models. Unsupervised models hold a substantial advantage of not requiring ground truth labels for model training, although model validation is more complicated. Populations with known (medication-induced) clinical fluctuations, or subjective symptom reports via ESM can be considered to test unsupervised models. The discriminative potential of the coefficient of variation found in chapter 9, might also contribute to the development of unsupervised models.

Last but not least, to enable reproducible findings, open-source, non-proprietary algorithms are needed [27,31]. The scientific community is doing a big effort since the last years to overcome this challenge [38,39], which should be continued and expanded.

10.4 Naturalistic motor symptom monitoring and adaptive DBS

Current status of adaptive DBS

Since the publication of chapter 1 essential work has been published regarding aDBS [40]. Following the demonstration of safety and tolerability of a fully implanted aDBS based on neurophysiological signals (neural-based aDBS) are demonstrated [41], the clinical effectivity has been demonstrated in long-term real-life experiments [42].

A better understanding of beta-oscillations suggested that aDBS working mechanisms rely on selective alteration of beta-distributions and beta-durations, rather than merely altering the total beta power [43–46]. Also, the concept of combining neural signals from the basal ganglia and the cortical motor area to improve neural aDBS [47] has been confirmed to contribute to naturalistic motor state decoding [42]. Furthermore, aDBS-paradigm differentiation is suggested between slow and fast algorithms [47], between different phenotypes [48–50], and between patients based on individual motor-state-specific neural signatures [42].

Naturalistic motor symptom monitoring to evaluate aDBS

Motion sensing is a strong candidate to assist in the evaluation of aDBS. Current aDBS literature suggests motion sensing to objectify acute motor effects [41,51], and to monitor motor effects in real-life [42]. Given that aDBS is a dynamic therapy, we argue that motion sensing is especially interesting for real-life aDBS monitoring. As explained before, clinician-based observations are too labour intensive and questionnaire-based assessments lack the desired temporal resolution for real-life monitoring. Research investigating the short-term validity of naturalistic motor monitoring will determine the role of motion sensing for aDBS evaluation. For this, the following twofold question is essential: '*Does the minimal window of time required for valid motion sensing 1) match the window on which the fast/slow aDBS paradigm requires revaluation (e.g. 1, 5, or 60 minutes), and 2) capture the difference in symptom severity which is intended to treat with an aDBS adjustment*'? If so, wearable motion sensing can be considered as aDBS input signal (motion-based aDBS). If not, motion sensing can still be used to evaluate neural-aDBS. Since available naturalistic tremor detection seems feasible in a nearer-future than bradykinesia detection [30,33], and DBS algorithms for tremor will be more likely a 'ON/OFF-algorithm' (chapter 1), motion-based aDBS might be introduced first for tremor-dominant rather than akinetic-rigid PD patients [52].

Another interesting neuroscientific debate is whether neural-based aDBS and motion-based aDBS are the same therapy. Technical advances will improve our understanding of daily life neurophysiological behavior [53,54] and may lead to millisecond precise neural-based aDBS algorithms. Neural-based aDBS tailored and responsive to specific neurophysiological recordings, will per definition have a different mechanism of action than motion-based aDBS.

First, this does not imply superiority of one of the two aDBS variants, but that they may rely on different neural working mechanisms. That is, how does the targeted neural target-feature behave under optimal and suboptimal treated conditions? And how do bradykinesia- or tremor-related movement features based on motion sensing, behave under different DBS settings? For both variants of aDBS similar DBS-algorithm development strategies have to be considered, as discussed in chapter 1. Possibly both variants need different time windows to yield the effect of a change in DBS paradigm, or different step sizes in between DBS paradigm changes.

Second, this requires similar proof-of-concept steps for motion-based aDBS, as done for neural-based aDBS [41,42,55,56]. Since motion-based aDBS experiments are not depending on operating theatres and externalized leads, they have the potential to progress faster.

10.5 Lessons learned regarding naturalistic PD motor monitoring

The inclusion of wearable clinical decision support systems (CDSS) can make PD monitoring more representative for the naturalistic fluctuations in real-life. Subjective eDiaries (such as ESM) have the potential to be used as naturalistic, repetitive ground truth in Parkinson monitoring. Practical feasibility and conceptual proof-of-concept are described here. A further validation of the subjective answers is required before ESM can be applied to develop and validate other wearable monitoring tools.

Wearable motion sensor tools are very promising to contribute to PD care as long-term evaluation devices. To determine the minimal naturalistic short-term fluctuation these models can detect, follow up research with more individual data is required. This minimal length will be determinant whether dynamic closed-loop therapies for PD can be programmed based on wearable motion sensing.

References

1. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 195 (2019).
2. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med.* **13**, 1 (2015).
3. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
4. Vasey, B. *et al.* DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**, 186–187 (2021).
5. Williams, A. *et al.* Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. *Lancet Neurol* **9**, 581–91 (2010).
6. Jaggi, J. L. *et al.* Bilateral Stimulation of the Subthalamic Nucleus in Parkinson's Disease: Surgical Efficacy and Prediction of Outcome. *Stereotact. Funct. Neurosurg.* **82**, 104–114 (2004).
7. Frizon, L. A. *et al.* Quality of Life Improvement Following Deep Brain Parkinson's Disease: Development of a Prognostic Model. *Neurosurgery* (2018) doi:10.1093/neuros/nyy287.
8. Cavallieri, F. *et al.* Predictors of Long-Term Outcome of Subthalamic Stimulation in Parkinson Disease. *Ann. Neurol.* doi:10.1002/ana.25994.
9. Kleiner-Fisman, G. *et al.* Subthalamic nucleus deep brain stimulation: summary and meta-analysis of outcomes. *Mov Disord* **21 Suppl 14**, S290-304 (2006).
10. Schuepbach, W. M. M. *et al.* Quality of life predicts outcome of deep brain stimulation in early Parkinson disease. *Neurology* (2019) doi:10.1212/wnl.0000000000007037.
11. Staartjes, V. E. & Kernbach, J. M. Letter to the Editor Regarding "Investigating Risk Factors and Predicting Complications in Deep Brain Stimulation Surgery with Machine Learning Algorithms". *World Neurosurg.* **137**, 496 (2020).
12. Deuschl, G. *et al.* A randomized trial of deep-brain stimulation for Parkinson's disease. *N Engl J Med* **355**, 896–908 (2006).
13. Odekerken, V. J. *et al.* Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson's disease (NSTAPS study): a randomised controlled trial. *Lancet Neurol* **12**, 37–44 (2013).
14. Schuepbach, W. M. *et al.* Neurostimulation for Parkinson's disease with early motor complications. *N Engl J Med* **368**, 610–22 (2013).
15. Vitek, J. L. *et al.* Subthalamic nucleus deep brain stimulation with a multiple independent constant current-controlled device in Parkinson's disease (INTREPID): a multicentre, double-blind, randomised, sham-controlled study. *Lancet Neurol* **19**, 491–501 (2020).
16. Shah, N. H., Milstein, A. & Bagley Ph, D. S. Making Machine Learning Models Clinically Useful. *Jama* (2019) doi:10.1001/jama.2019.10306.
17. Cebi, I., Scholten, M., Gharabaghi, A. & Weiss, D. Clinical and Kinematic Correlates of Favorable Gait Outcomes From Subthalamic Stimulation. *Front. Neurol.* **11**, 212 (2020).
18. Younce, J. R., Campbell, M. C., Perlmutter, J. S. & Norris, S. A. Thalamic and ventricular volumes predict motor response to deep brain stimulation for Parkinson's disease. *Park. Relat Disord* (2018) doi:10.1016/j.parkreldis.2018.11.026.
19. Shang, R., He, L., Ma, X., Ma, Y. & Li, X. Connectome-Based Model Predicts Deep Brain Stimulation Outcome in Parkinson's Disease. *Front. Comput. Neurosci.* **14**, (2020).
20. Victor J. Geraedts. Right on Track: Towards improving DBS patient selection and care. (Leiden University, 2020).
21. Rizzone, M. G., Martone, T., Balestrino, R. & Lopiano, L. Genetic background and outcome of Deep Brain Stimulation in Parkinson's disease. *Parkinsonism Relat. Disord.* **64**, 8–19 (2019).
22. Pedersen, M. *et al.* Artificial intelligence for clinical decision support in neurology. *Brain Commun.* **2**, (2020).
23. Mateen, B. A., Liley, J., Denniston, A. K., Holmes, C. C. & Vollmer, S. J. Improving the quality of machine learning in health applications and clinical research. *Nat. Mach. Intell.* **2**, 554–556 (2020).
24. Pencina, M. J., Goldstein, B. A. & D'Agostino, R. B. Prediction Models - Development, Evaluation, and Clinical Application. *N Engl J Med* **382**, 1583–1586 (2020).
25. Jakobs, M., Fomenko, A., Lozano, A. M. & Kiening, K. L. Cellular, molecular, and clinical mechanisms of action of deep brain stimulation—a systematic review on established indications and outlook on future developments. *EMBO Mol. Med.* **11**, e9575 (2019).

26. Sullivan, I., DeHaven, A. & Mellor, D. Open and Reproducible Research on Open Science Framework. *Curr. Protoc. Essent. Lab. Tech.* **18**, e32 (2019).
27. Espay, A. J. *et al.* A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Mov Disord* **34**, 657–663 (2019).
28. Todorova, A., Jenner, P. & Ray Chaudhuri, K. Non-motor Parkinson's: integral to motor Parkinson's, yet often neglected. *Pract. Neurol.* **14**, 310 (2014).
29. Habets, J. & Kubben, P. EMA and wearable sensor monitoring in PD. (2020) doi:10.34894/5HHK8H.
30. Thorp, J. E., Adamczyk, P. G., Ploeg, H. L. & Pickett, K. A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. *Front Neurol* **9**, 1036 (2018).
31. Warmerdam, E. *et al.* Long-term unsupervised mobility assessment in movement disorders. *Lancet Neurol.* **19**, 462–470 (2020).
32. Fasano, A. & Mancini, M. Wearable-based mobility monitoring: the long road ahead. *Lancet Neurol.* **19**, 378–379 (2020).
33. Powers, R. *et al.* Smartwatch inertial sensors continuously monitor real-world motor fluctuations in Parkinson's disease. *Sci. Transl. Med.* **13**, eabd7865 (2021).
34. Rodriguez-Molinero, A. Monitoring of Mobility of Parkinson's Patients for Therapeutic Purposes - Clinical Trial (MoMoPa-EC). NCT04176302 (2019).
35. Ossig, C. *et al.* Correlation of Quantitative Motor State Assessment Using a Kinetograph and Patient Diaries in Advanced PD: Data from an Observational Study. *PLoS One* **11**, e0161559 (2016).
36. Lonini, L. *et al.* Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *Npj Digit. Med.* **1**, 64 (2018).
37. Mahadevan, N. *et al.* Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *NPJ Digit Med* **3**, 5 (2020).
38. Rochester, L. *et al.* A Roadmap to Inform Development, Validation and Approval of Digital Mobility Outcomes: The Mobilise-D Approach. *Digit. Biomark.* **4**, 13–27 (2020).
39. Kluge, F. *et al.* Consensus based framework for digital mobility monitoring. *medRxiv* 2020.12.18.20248404 (2020) doi:10.1101/2020.12.18.20248404.
40. Bouthour, W. *et al.* Biomarkers for closed-loop deep brain stimulation in Parkinson disease and beyond. *Nat. Rev. Neurol.* **15**, 343–352 (2019).
41. Velisar, A. *et al.* Dual threshold neural closed loop deep brain stimulation in Parkinson disease patients. *Brain Stimulat.* **12**, 868–876 (2019).
42. Gilron, R. *et al.* Long-term wireless streaming of neural recordings for circuit discovery and adaptive stimulation in individuals with Parkinson's disease. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00897-5.
43. Tinkhauser, G. *et al.* The modulatory effect of adaptive deep brain stimulation on beta bursts in Parkinson's disease. *Brain* **140**, 1053–1067 (2017).
44. Deffains, M., Iskhakova, L., Katabi, S., Israel, Z. & Bergman, H. Longer β oscillatory episodes reliably identify pathological subthalamic activity in Parkinsonism. *Mov. Disord.* **33**, 1609–1618 (2018).
45. M. N. Petrucci *et al.* A Closed-loop Deep Brain Stimulation Approach for Mitigating Burst Durations in People with Parkinson's Disease. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 3617–3620 (2020). doi:10.1109/EMBC44109.2020.9176196.
46. Duchet, B. *et al.* Average beta burst duration profiles provide a signature of dynamical changes between the ON and OFF medication states in Parkinson's disease. *bioRxiv* 2020.04.27.064246 (2021) doi:10.1101/2020.04.27.064246.
47. Little, S. & Brown, P. Debugging Adaptive Deep Brain Stimulation for Parkinson's Disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **35**, 555–561 (2020).
48. Canessa, A., Palmisano, C., Isaias, I. U. & Mazzoni, A. Gait-related frequency modulation of beta oscillatory activity in the subthalamic nucleus of parkinsonian patients. *Brain Stimulat.* **13**, 1743–1752 (2020).
49. Petrucci, M. N. *et al.* Neural closed-loop deep brain stimulation for freezing of gait. *Brain Stimulat.* **13**, 1320–1322 (2020).
50. Godinho, F. *et al.* Spectral Characteristics of subthalamic Nucleus local Field Potentials in parkinson's disease: Phenotype and Movement matter. *Eur. J. Neurosci.* **n/a**, (2021).
51. Castaño-Candamil, S. *et al.* A Pilot Study on Data-Driven Adaptive Deep Brain Stimulation in Chronically Implanted Essential Tremor Patients. *Front. Hum. Neurosci.* **14**, (2020).
52. Malekmohammadi, M. *et al.* Kinematic Adaptive Deep Brain Stimulation for Resting Tremor in Parkinson's Disease. *Mov Disord* **31**, 426–8 (2016).

53. Thenaisie, Y. *et al.* Towards adaptive deep brain stimulation: clinical and technical notes on a novel commercial device for chronic brain sensing. *medRxiv* 2021.03.10.21251638 (2021) doi:10.1101/2021.03.10.21251638.
54. Feldmann, L. K. *et al.* Subthalamic beta band suppression reflects effective neuromodulation in chronic recordings. *Eur. J. Neurol.* **n/a**,.
55. Little, S. *et al.* Adaptive deep brain stimulation for Parkinson's disease demonstrates reduced speech side effects compared to conventional stimulation in the acute setting. *J. Neurol. Neurosurg. Psychiatry* **87**, 1388–1389 (2016).
56. Little, S. *et al.* Adaptive deep brain stimulation in advanced Parkinson disease. *Ann Neurol* **74**, 449–57 (2013).

Chapter 11

Summary

Parkinson's disease (PD) is the second most common neurodegenerative disorder influencing the life of millions of people worldwide. It is characterized by motor and non-motor symptoms which worsen over time and affect the daily quality of life of patients. 5% to 10% of PD patients will be considered for deep brain stimulation (DBS), when pharmacologic dopamine-replacement therapy is not alleviating the motor symptoms to a satisfactory level (anymore). The aim of this thesis was to improve future DBS care for PD patients via individual preoperative outcome prediction and real-life (naturalistic) symptom monitoring.

The work in part A focused on improving postoperative DBS outcome in PD by improving individual outcome prediction preoperatively. We hypothesized that a better recognition of patient who are likely to have a suboptimal response to DBS, may lead to improved patient counselling and potentially patient selection. Chapter 2 demonstrated the principle of predicting individual DBS outcomes with a machine learning model which analysed preoperative clinical variables. The presented model was able to discriminate between strong and weak responders with an area under the curve of 0.79 (standard deviation (sd): 0.08) and a classification accuracy of 78%. In chapter 3, we validated a generalizable version of this model in an external cohort of 322 PD patients who underwent subthalamic nucleus DBS over six international DBS centres. The model was able to discriminate between strong and weak responders with an area under the curve of 0.76 and a classification accuracy up to 77%.
During the design, execution, and evaluation of these studies, we encountered some inherent challenges of preoperative individual outcome prediction in DBS practice. We highlighted essential steps to realize actual clinical impact for the patient, after the successful retrospective development and validation of a model. Binary, or standardized regressive, outcome definition is very challenging in DBS care, and a proposed outcome should be validated to some extent compared to traditional gold standards for DBS success. Also, the interaction between clinician, patient, and prediction tool must be studied to evaluate a model's potential therapeutic and socioeconomical effect. The work in part A of this thesis can be part of the foundation of future clinical decision support systems for DBS care.

In part B, we focused on the real-life (naturalistic) monitoring of PD symptoms. We were especially interested in the potential of naturalistic monitoring to evaluate dynamic, self-adjusting DBS therapies (adaptive DBS). First, we demonstrated how a smartphone-based self-reporting diary (experience sampling method, ESM), combined with wearable motion sensors, can contribute to naturalistic Parkinson monitoring. We studied 20 PD patients who monitored themselves with ESM (8 identical questionnaires on motor, non-motor, and context per day) and wearable motion sensors for two weeks. In chapter 5 we demonstrated feasibility of this combined method since 79% of ESM-questionnaires were completed, and the wearable sensors successfully collected motion data during 94% of the time. In chapter 6 we concluded that the subjective answers reporting on, among others, motor symptoms, non-motor symptoms, and motor functionality, were coherent and valid and could be useful in naturalistic PD monitoring.
Since the long-term application of ESM is new in PD monitoring, and the combined monitoring methodology of ESM and wearable motion sensors was not described before, we made the total data set publicly available. To motivate and help other researchers to use these combined data types, we described our methodology of combining and analysing the two data types in chapter 7.

In chapter 8 we demonstrated our suggested application to combine ESM and wearable sensors to monitor naturalistic motor symptoms. In a single PD patient suffering from severe tremor fluctuations, we used the ESM answers as ground truth to train a machine learning model to predict tremor severity. Analysing windows of 15 minutes of motion data, the model predicted tremor severity with a significant correlation of $r = 0.43$.

In chapter 9, we investigated on which short-term time window, motion sensors were able to differentiate naturalistic medication-induced motor fluctuations. In 18 out of 20 PD patients, differentiation of (bradykinesia-centred) motor fluctuations was feasible on a one-minute window and performance increased when five-minute windows were analysed. Our hypothesis that models trained on individual data would outperform models trained on group data could not be confirmed. However, a replication of these analyses with larger individual data sets is necessary to conclude whether there is no additional value of individual model training.

Concluding, in part A we demonstrated and validated the potential of individual DBS outcome prediction to add clinical patient value, and highlighted which inherent challenges must be overcome to fulfil this potential. In part B, we applied ESM as a method to improve naturalistic PD monitoring and provided a proof of concept how this can practically improve naturalistic motor symptom monitoring. We described the feasibility of motor fluctuation differentiation on a one-to-five-minute level, and provided future steps to investigate whether motion sensing can be used to control dynamic therapies as adaptive DBS.

Chapter 12

Impact and valorization

Both the findings on individual preoperative outcome prediction for deep brain stimulation (DBS) in Parkinson's disease (PD), as on naturalistic monitoring in PD belong to the early development stages of potential future clinical decision support systems (CDSS). For both topics, the answers on the hypotheses enable next scientific steps towards CDSS which should optimize patient care in PD.

The described individual preoperative outcome prediction can improve patient counselling and maybe even patient selection in the future. This may lead to a positive socioeconomic impact. First, prospective follow up research should be done to optimize the target population and the predictive model, to validate predictive performance in a prospective setting, and to investigate the interaction between CDSS, clinician, and patient. A valorising partner such as a neuro- or health care-tech start-up may be included to enable a smoother transition from the scientific setting into an applicable CDSS. The lessons presented in our discussion are likely to be beneficial for researchers and clinicians considering CDSS for other neurological and psychiatric DBS-indications.

The proven long-term feasibility of electronic diaries (such as experience sampling method (ESM)) can contribute to more usage of ESM in clinical practice, and to a better understanding between patient and clinician about the naturalistic symptom fluctuations. Further, we encouraged the scientific community to combine ESM and motion sensors to improve naturalistic PD motor monitoring.

The demonstrated feasibility and potential methodologies of short-term naturalistic bradykinesia fluctuation detection may especially improve the evaluation and management of PD motor therapies. All this can improve real-life symptom assessment, provide the clinician with a more accurate therapeutic evaluation, and lead to better tailored therapeutic management in PD.

Appendices

Nederlandse samenvatting

De ziekte van Parkinson (ZvP) is de tweede meest voorkomende neurodegeneratieve aandoening en beïnvloedt miljoenen mensen wereldwijd. Kenmerkend voor de ZvP zijn toenemende motorische en niet-motorische klachten die de kwaliteit van leven verminderen. 5% tot 10% van de patiënten lijdend aan de ZvP komt in aanmerking voor diepe hersenstimulatie (DBS) wanneer farmacologische behandeling geen afdoende effect meer heeft op de motorische symptomen. Het doel van dit proefschrift is om de DBS zorg voor de ZvP te verbeteren middels preoperatieve uitkomst voorspelling en symptoom monitoring in het dagelijks leven van patiënten.

Het werk in deel A richt zich op het verbeteren van postoperatieve resultaten na DBS door de voorspelling van de uitkomst preoperatief te verbeteren. Onze hypothese was patiënt voorlichting en wellicht patiënt selectie verbeterd kan worden door een betere inschatting van de individuele kans op een gunstige uitkomst na DBS. In hoofdstuk 2 toonden we het principe van individuele uitkomst voorspelling middels een machine learning model op basis van preoperatieve klinische variabelen. Het model differentieerde tussen patiënten met een goede en slechte uitkomst met een area under the curve (AUC) van 0.79 (standaarddeviatie 0.08) en een classificatie accuracy van 78%. In hoofdstuk drie hebben we dit model (na minimale aanpassing van de generaliseerbaarheid) getest in een extern patiëntcohort van 322 patiënten die DBS ondergingen voor de ZvP uit zes internationale DBS-centra. Het model differentieerde de patiënten met goede en slecht uitkomst met een AUC van 0.76 en een accuracy tot 77%.
Tijdens de ontwikkeling, uitvoering en evaluatie van deze studies zijn we meerdere uitdagingen tegengekomen die inherent zijn aan preoperatieve uitkomst voorspelling voor DBS. We hebben essentiële stappen beschreven die nodig zijn om de patiëntenzorg daadwerkelijk te verbeteren. Ten eerste is een gestandaardiseerde DBS-uitkomstdefinitie ingewikkeld op te stellen. Elke voorgestelde definitie zal tot zekere hoogte moeten worden gevalideerd met erkende uitkomstmaten. Ook moet de interactie tussen de behandelaar, de patiënt en het model worden onderzocht om een beeld te krijgen van het werkelijke therapeutische en socio-economische effect. Het werk in deel A kan deel van de basis zijn voor toekomstige klinische beslissing-ondersteunende modellen in de DBS-zorg.

In deel B hebben we ons gericht op het monitoren van de ZvP in het dagelijks leven. In het bijzonder waren we geïnteresseerd of dagelijks-leven-monitoring zelfsturende therapieën zoals adaptieve DBS kan aansturen. Eerst hebben we aangetoond hoe een smartphone-vragenlijst (experience sampling method, ESM) gecombineerd met bewegingssensoren de dagelijks-level-monitoring kunnen verbeteren. Hiervoor bestudeerden we 20 patiënten met de ZvP die gemonitord werden met ESM (8 identieke vragenlijsten over hun motorische en niet-motorische klachten en hun omgeving per dag) en bewegingssensoren gedurende twee weken. Hoofdstuk 5 beschrijft de haalbaarheid van deze combinatie gezien 79% van de ESM-vragenlijsten werd beantwoord en 94% van de beoogde tijd was geregistreerd door de bewegingssensoren. Hoofdstuk 6 beschrijft dat de ingevulde ESM-antwoorden wat betreft onder andere motorische en niet-motorische symptomen en bewegingsfunctionaliteit een logische samenhang vertoonden. Dit spreekt er voor dat ESM kan worden toegepast in het monitoren van ZvP-klachten in het dagelijks leven.
Aangezien zowel dit langdurige gebruik van ESM bij de ZvP, als ook de combinatie met bewegingssensoren, nog niet beschreven was, hebben we de volledige data publiekelijk

beschikbaar gemaakt. In hoofdstuk 7 beschrijven we ons methode om de verschillende data types te combineren en te analyseren om andere onderzoekers te helpen en motiveren.
In hoofdstuk 8 tonen we hoe de voorgestelde combinatie van ESM en bewegingssensoren kan bijdragen aan het monitoren van motorische symptomen in het dagelijks leven. Bij één patiënt met forse tremor fluctuaties hebben we de ESM-antwoorden gebruikt als goud standaard om een model te ontwikkelen dat de mate van tremor herkent aan de bewegingssensoren registraties. Wanneer blokken van 15 minuten bewegingsdata werden gebruikt om de tremor scores te voorspellen, kwamen deze overeen met de daadwerkelijk scores met een correlatie van r = 0.43.

In hoofdstuk 9 hebben we specifiek gekeken of bewegingssensoren de motorische fluctuaties in het dagelijks leven op korte tijdsbestekken kunnen herkennen. Het herkennen van motor fluctuaties (gericht op bradykinesie) per minuut was mogelijk in 18 van 20 patiënten met de ZvP. De voorspellingsprestaties verbeterden wanneer er werd gekeken per vijf minuten. We konden onze hypothese dat modellen ontwikkeld met individuele data secuurder zijn dan modellen ontwikkeld op groepsdata niet bevestigen. Herhaling van deze analyses met langere individuele datacollecties is echter aangewezen voordat er over de toegevoegde waarde van individueel getrainde modellen kan worden geconcludeerd.

Samenvattend hebben we in deel A laten zien dat individuele uitkomstvoorspelling mogelijkheden biedt om de zorg voor patiënten met de ZvP die in aanmerking komen voor DBS te verbeteren. We hebben nieuwe validatiestappen gezet en besproken wat de volgende onderzoek stappen moeten zijn. In deel B hebben we laten zien hoe ESM een toegevoegde waarde kan zijn om de monitoring te verbeteren van motorische symptomen in het dagelijks leven van patiënten met de ZvP. Daarnaast lieten we zien dat het detecteren van motorische fluctuaties in het dagelijks leven ook mogelijk is op tijdsbestekken van een tot vijf minuten, en hebben we methodes aangereikt waarmee onderzocht kan worden of deze monitoring secuur genoeg kan worden om adaptieve DBS mee aan te sturen.

Dankwoord

Prof. Temel, beste Yasin, ik wil je bedanken voor het vertrouwen en de kansen die je mij gegeven hebt. Ik bewonder hoe je enerzijds steeds weer binnen enkele minuten de vinger op de zere wetenschappelijke plek weet te leggen, en anderzijds nog tijd weet te maken voor het voetbal van je meiden!

Dr. Kubben, beste Pieter, ik wil je graag bedanken voor de dagelijkse begeleiding en de fijne sfeer. Ik waardeer het hoe je altijd nieuwe ideeën wilde ondersteunen (als het even kon). Dankzij jou werden Yasin's woorden over de 'mooie PhD-jaren' inderdaad waarheid!

Dr. Kuijf, beste Mark, ik wil je graag bedanken voor alle begeleiding en waardevolle inzichten. Ondanks dat het contact niet wekelijks hoefde te zijn, was je altijd bereikbaar voor waardevolle toevoegingen.

Dr. Herff, lieber Christian, this thesis wouldn't have been the same without you joining the team during my first year. Thank you for supervising on my first data science steps and for being an exemplary daily supervisor, and last but not least for being a wonderful colleague!

Dr. Janssen, beste Mark, jij was waarschijnlijk de eerste die mij enthousiast maakte over onderzoekstijd, dankjewel voor alle succesvolle én onsuccesvolle projecten en vriendschap, hopelijk tot snel in Berlijn!

Dr(s). Heijmans, beste Margot, ik ben super blij dat we de afgelopen 4 jaar samen doorlopen hebben en altijd bij elkaar terecht konden. Dankjewel voor alle gezelligheid en heel veel success met jouw volgende stappen!

Beste Mirella, Nicoles (P + B), Linda, Annelien, Simone (en ESM expert-groep) en Albert, dank jullie wel voor alle goede adviezen die jullie mij afgelopen vier jaar gegeven hebben.

Prof. Starr, dr. Little, dr. Gilron and mr. Wilt, dear Phil, Simon, Ro'ee and Bobby, thank you for a truly amazing and inspiring time in San Francisco. It was a privilege to work with you. Thanks for the expertise and friendship, and I hope we keep up the contact.

Dr. Beudel, beste Martijn, bedankt voor de prettige samenwerking over de jaren, het was altijd makkelijk aankloppen en nooit teleurstellend jou ergens voor te vragen.

Beste stafleden en A(N)IOS Neurochirurgie MUMC+, bedankt voor de leuke en leerzame jaren die mij gebracht hebben waar ik nu ben. Maastricht verlaten was een moeilijke beslissing, maar ik hoop dat onze wegen in de toekomst blijven kruisen. In het bijzonder bedankt Mariel, ik blijf je voor altijd dankbaar voor de eerlijke raad over mijn stotter.

Dear UNS50-roommates, Liu, Maarten, Majed, Margot and Roman, thanks for being such nice roommates and all shared coffee, chocolate, gezelligheid and awkwardness.
Thanks to all (former) Temel Lab-colleagues for all coffee, BBQ's, dinners; and thanks to all other division-3 colleagues for all shared lunch, coffee, beers and complaining.

Thanks to all Brainballers for all precious soccer moments!

Dear McGuire Program family, thanks to everyone who is working hard every day to become a strong speaker. It is so inspiring to see everyone's dedication to improve themselves and the community surrounding them.

Dear Alex, thanks for being our SF-lockdown-friend, hope to see you again!

Tommie, Kuun en Leon, bedankt voor het ontwerpen van de prachtige kaft van dit boekje.

Dank aan alle vrienden- en niet-vrienden-groepen voor de broodnodige fysieke en mentale afleiding de afgelopen 4 jaar. Zonder jullie was dit boekje er ook niet geweest.
JDB/Stamppot-Club, Merci; Zvv Hazendans 6, bedankt voor de kleedkamer-pret, niet-gedekte spelers en goals; Stookers, voor de grappa; NoLimit, voor de culturele stedentrips; Abio, voor alle spontane festiviteiten; Lowlands-crew, voor de legendarische momenten in en rond de X-Ray! Tenslotte Vic en Bern, Kas en Lucas, Chip, voor alle Rotterdamse gezelligheid.

Liebe BeHouse Family, thanks for being my second family in Berlin. All the shared food, stories, games, coffee, and of course parties made home office a lot easier the last year.

Koen, merci voor het dagelijks delen van lief en leed de afgelopen jaren.
Dear Soof, Isa and Manolis, thanks for the warm CVC vibes, and hope to see you again all over Europe. Special thanks to Terry for sharing our office. Special thanks to Maikel for completing the furniture.

Lieve Sjoerd, bedankt voor onze vriendschap in lief en leed, en door dik en dun(!). Die uren aan de telefoon zijn mij erg dierbaar. Snel tijd met jou, Marleen en Fientje door Berlijn te struinen.

Lieve Camiel, dankjewel voor de band die we hebben en dat we alles met elkaar kunnen bespreken. Merci voor de broer en vriend die je bent!

Lieve Papa, dankjewel voor alle liefde en ondersteuning de afgelopen meer dan 30 jaar. Zonder al die energie was dit boekje er ook nooit geweest. Ik ben trots op onze band die de afgelopen jaren alleen maar beter is geworden.

Lieve Mama, hoe pijnlijk het is dat ik dit moment niet met je kan delen, zo mooi is het dat je me elke dag weer inspireert, soms begin ik je alleen maar beter te begrijpen naarmate de tijd verstrijkt. Bedankt voor alles.

Liebste Maha, I am so grateful that we walked into each other's life, that night am Schluchsee became lifechanging. The love, fun, food, activities, and conversations we can share makes me so proud and confident about our plans. Can't thank you enough for being at my side.

Curriculum vitae

Jeroen Habets was born on November 27th 1989 in Sittard, the Netherlands. In 2008, he graduated from pre-university education Gymnasium at the Porta Mosana College in Maastricht.
In 2009, he started his Medicine study at Maastricht University, after studying Business Engineering for one year at the Technical University Eindhoven. During the medical bachelor studies, he completed the Honors Program International Health with a mobility and thesis about health care structures in rural South-India. Next to his studies, he did several volunteering and organization activities for the student-soccer association Red Socks and founded a hobby distillery with several friends.
After graduation in 2015, he worked as a resident (not in training) at the department of Neurosurgery at Maastricht University Medical Center. Next to his clinical and scientific activities, he was part of the board and chairman of the residents association (AAV) of Maastricht UMC+, and part of the organizing committee of the International Conference on Congenital Hydropcephalus, in Maastricht.

In September 2017, he started the work which led to this thesis as a full time PhD-candidate, under supervision of Professor Temel, Dr. Kubben and Dr. Kuijf, and from 2018 on Dr. Herff. During these years, he completed several online courses and a summer school on programming in Python, signal processing, and machine learning.
The work in chapter 4 got awarded with the prize for best poster presentation at the International Clinical Neuroscience Course 2017 in Samsun, Turkey, and the work in chapter 3 got awarded with the prize for best abstract and oral presentation at the International Course on Neuromodulation 2019 in Maastricht, The Netherlands.
He received several personal travel grants from ZonMW, the Dutch Parkinson association, and the Young European Research Universities Network to conduct parts of the work in this thesis at the department of Neurological Surgery at the University of California San Francisco (UCSF) during four months in 2020, supervised by Professor Starr, Dr. Little and Dr. Gilron, and at the Cognitive Sciences Lab at the University of Bremen in 2018 during two weeks supervised by Professor Schultz and Dr. Herff.

Jeroen moved to Berlin in Summer 2021 where he will continue his medical and scientific career.

List of publications

"A long-term, real-life Parkinson monitoring database combining unscripted objective and subjective recordings"
Habets, Heijmans, Simons, Leentjens, Temel, Kuijf, Kubben, Herff.
Data, 2021, 6(2) 22.

"Mobile Health Daily Life Monitoring for Parkinson Disease: Development and Validation of Ecological Momentary Assessments"
Habets, Heijmans, Herff, Simons, Leentjens, Temel, Kuijf, Kubben.
JMIR mHealth and uHealth, 2020, 8(5), e15628.

"Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease—proof of principle in a retrospective cohort"
Habets, Janssen, Duits, Sijben, Mulders, De Greef, Temel, Kuijf, Kubben, Herff.
PeerJ, 2020, 8, e1031.

"Future Perspectives: Adaptive Deep Brain Stimulation",
Beudel, Heijmans, **Habets**, Kubben.
In book: Fundamentals and Clinics of Deep Brain Stimulation.
Editors: Temel, De Bie, Chabardes, Fasano. Springer Nature, 2020.

"Monitoring Parkinson's disease symptoms during daily life: A feasibility study"
Heijmans, **Habets**, Herff, Aarts, Stevens, Kuijf, and Kubben.
npj Pakrinson's Disease, 2019, 5(21).

"Evaluation of Parkinson's Disease at Home: Predicting Tremor from Wearable Sensors"
Heijmans, **Habets**, Kuijf, Kubben, Herff;
Full contributed 4-page paper at the 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Berlin, Germany, July 2019

"Controversies in Deep Brain Stimulation Surgery: Micro-Electrode Recordings",
Habets, Isaacs, Vinke, Kubben.
In book: Evidence for Neurosurgery, Effective Procedures and Treatment. Editors: Bartels, Rovers, Westert. Springer Nature, 2019.

"An update on adaptive deep brain stimulation in Parkinson's disease"
Habets & Heijmans, Kuijf, Janssen, Temel, Kubben.
Mov Disorders 2018, Dec;33(12):1834-1843.

"Letter to the editor on "A rare case of Candida glabrata spondylodiscitis: Case report and literature review" by Gagliano et al."
Habets, Teernstra, van Hemert, Linssen, Heijboer, Rijkers.
Int J Infect Diseases 2018, Jul;72:19.

“Acute Monocular Blindness Due to Orbital Compartment Syndrome Following Pterional Craniotomy”
Habets, Haeren, Lie, Bauer, Dings.
World Neurosurgery, 2018 Jun; 114:72-75.

“Serious and reversible levetiracetam-induced psychiatric symptoms after resection of frontal low grade glioma: two case histories”
Habets, Leentjens, Schijns.
British Journal of Neurosurgery, 2017 Aug;31(4):471-473.

"Anterior transthoracic surgery with MEP monitoring support for high-risk thoracic disc herniations"
Cornips, **Habets**, Van Kranen – Mastenbroek, Bergs, Bos, Jacobi –Postma.
World Neurosurgery, 2017 Sep;105:441-455.