# Smartphone Speech Testing for Symptom Assessment in Rapid Eye Movement Sleep Behavior Disorder and Parkinson's Disease

**Siddharth Arora[1,2,3], Christine Lo[4,5], Michele Hu[4], and Athanasios Tsanas[6], Senior Member, IEEE**

[1]Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK
[2]Somerville College, University of Oxford, Oxford, OX2 6HD, UK
[3]Saïd Business School, University of Oxford, Oxford, OX1 1HP, UK
[4]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, UK
[5]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, OX3 9DU, UK
[6]Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, EH16 4UX, UK

Corresponding author: S. Arora (arora@maths.ox.ac.uk)

The research was funded by the Monument Trust Discovery Award from Parkinson's UK (J-1403). Christine Lo was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the National Health Service (NHS), the NIHR or the Department of Health.

**ABSTRACT** Speech impairment in Parkinson's Disease (PD) has been extensively studied. Our understanding of speech in people who are at an increased risk of developing PD is, however, rather limited. It is known that isolated Rapid Eye Movement (REM) sleep Behavior Disorder (RBD) is associated with a high risk of developing PD. The aim of this study is to investigate smartphone speech testing to: (1) distinguish participants with RBD from controls and PD, and (2) predict a range of self- or researcher-administered clinical scores that quantify participants' motor symptoms, cognition, daytime sleepiness, depression, and the overall state of health. The rationale of our analyses is to test an initial hypothesis that speech can be used to detect and quantify the symptoms associated with RBD and PD. We analyzed 4242 smartphone voice recordings collected in clinic and at home from 92 Controls, 112 RBD and 335 PD participants. We used acoustic signal analysis and machine learning, employing 337 features that quantify different properties of speech impairment. Using a leave-one-subject-out cross-validation scheme, we were able to distinguish RBD from controls (sensitivity 60.7%, specificity 69.6%) and RBD from PD participants (sensitivity 74.9%, specificity 73.2%), and predict clinical assessments with clinically useful accuracy. These promising findings warrant further investigation in using speech as a digital biomarker for PD and RBD to facilitate intervention in the early and prodromal stages of PD.

**INDEX TERMS** Digital biomarkers, Parkinson's disease, REM sleep behavior disorder, speech analysis, statistical learning, smartphones, telemedicine.

## I. INTRODUCTION

Neurological disorders pose an increasing burden to health systems worldwide as leading sources of disability [1]. Parkinson's Disease (PD) is characterized by a range of progressively debilitating motor symptoms (including bradykinesia, tremor, rigidity) and non-motor (e.g. cognitive, neuropsychiatric, autonomic, sleep) symptoms [2]. Speech performance degradation is reported in the vast majority of people diagnosed with PD and speech-related problems are strongly associated with overall PD symptom severity [3], [4].

There is currently no known cure for PD, however, pharmacological and surgical treatment can to some extent alleviate the symptoms and improve quality of life for most People with PD (PWP) [5]. Regular monitoring of symptom progression is indicated to optimize treatment regimens, which has relied on expert-based clinical assessments and PWP's self-reports. Clinical assessment relies on established validated instruments (rating scales). One of the most widely used instruments is the Movement Disorder Society (MDS)-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [2], which requires skilled expert raters to administer. Even though this is well standardized and rater training is administered, similarly to other expert-rated scales the MDS-UPDRS is known to be prone to inter-rater

variability [6]. Additionally, the time required to administer the MDS-UPDRS often prohibits its routine clinical use.

Expert administered clinical assessments provide a clinical impression of symptom severity and are well-suited for non-motor and motor-tasks that are more amenable to objective external assessment. However, it is crucial to consider the PWP's self-perception of symptom severity, since ultimately different PWP have different needs [1]. The proliferation of smartphones and smartphone apps has facilitated longitudinal collection of Patient Reported Outcome Measures (PROMs) for participant symptom self-reporting [7]–[9]. This further motivates the need to use PROMs for PD management and monitoring of the diverse range of PD symptoms.

In addition to clinically validated rating scales (expert-based assessments and PROMs), the research community has embraced the use of technology in the hope of facilitating objective, sensor-based PD assessments [10]. These developments include the use of wearable sensors [11] and smartphones [12]–[14]. We have previously demonstrated the use of sustained vowel "aaah" towards: (i) very accurate binary differentiation of a matched control group versus PWP [15]; (ii) replication of MDS-UPDRS with greater accuracy than the inter-rater variability [3], [16]–[18]; (iii) automatic assessment of PD voice rehabilitation using the Lee Silverman Voice Treatment (LSVT) [19]; (iv) distinguishing people with genetic PD predisposition (Leucine-Rich Repeat Kinase 2 (LRRK2) mutations [20]), PWP, and matched controls; and (v) using a large database of voice recordings collected over a standard telephone network (sampling frequency 8 KHz) to distinguish PWP from age- and sex-matched controls [21].

Voice abnormalities have been reported to precede the onset of motor symptoms in PD [22]. Investigating the nature and extent of vocal impairment in individuals who are at risk of developing PD can provide a crucial opportunity to intervene in the prodromal stages of the disease and facilitate potential recruitment of participants for neuroprotective treatment trials aimed at slowing down or preventing conversion to PD. Rapid Eye Movement (REM) sleep Behavior Disorder (RBD) is among the strongest known predictors of PD risk [23]. Isolated RBD is associated with high rates of phenoconversion to a neurodegenerative disorder, including PD, dementia with Lewy bodies, and multiple system atrophy [24]. Isolated RBD is a parasomnia that is typically characterized by dream enactments and excess muscle tone during REM sleep [25]–[27]. Age and sex are the two most common risk factors associated with RBD, whereby there is a higher preponderance in males. The risk of developing a neurodegenerative syndrome, from the time of RBD diagnosis, is estimated to be 33.1 at five years, 75.5% at ten years, and 90.9% at 14 years, with a median conversion time of 7.5 years [24]. The aforementioned reasons motivated our decision to investigate the signs and extent of potential vocal impairment in participants with isolated RBD.

For the assessment of RBD, a screening questionnaire is sometimes employed (REM Sleep Behavior Disorder Screening Questionnaire (RBDSQ)), and the gold standard for RBD diagnosis is a polysomnography (PSG) test. Administering full PSG incurs substantial logistical costs for the healthcare service providers (as the participants need to be admitted and monitored throughout the night at hospital). The average cost of an overnight PSG test is estimated to be around USD 800 [28]. Voice analysis offers the exciting possibility to risk stratify individuals and prioritize those who are most likely to benefit from a PSG investigation.

The literature on investigating vocal impairment in individuals with RBD is rather scarce. Using speech recordings (sustained phonation, syllable repetition and monologue) from 16 RBD and 16 age- and sex-matched healthy controls, a sensitivity of 96% and specificity of 79% has been reported [29]. Using 50 RBD, 30 PD and 30 healthy controls, an Area Under the Curve (AUC) value of 0.69 (sensitivity 69.8%, specificity 64.7%) was achieved in discriminating RBD and controls using smartphone-based speech, and a high correlation and reliability were found between acoustic measures extracted from a professional microphone and smartphone [30]. These findings suggest that recordings collected from smartphones and professional microphones could be of comparable quality. Using the speech dataset employed by Rusz et al. [30], a classification of up to 66% between early PD and RBD was reported by Benba et al. [31]. However, these studies on speech-RBD have mainly relied on high-quality recordings, collected in a laboratory under controlled acoustic conditions, using small cohorts (typically fewer than 50 participants). Thus, current studies may be rather limited in drawing inferences and scaling findings for screening people with isolated RBD. Moreover, in the absence of detailed clinical measures of key interest, studies thus far have been unable to offer new insights into the relationship between the extent of speech impairment and severity of symptoms in RBD.

The aim of this study is to utilize smartphone speech assessments to make the following three main contributions: (1) differentiating cohorts of healthy controls ($n$ = 92), isolated RBD ($n$ = 112), and PD ($n$ = 335) participants using sustained vowel phonations; (2) predicting diverse self- or researcher-rated established validated clinical metrics assessing symptom severity from a deeply phenotyped cohort; (3) highlighting the benefits of deep clinical phenotyping to fully maximize the application of smartphone speech evaluation for RBD and PD. We aim to provide an overview of symptoms in RBD and early PD by bridging objective data collected using smartphones (voice), clinical ratings (e.g., MDS-UPDRS), and self-reports, with the ultimate aim of contributing towards the development of a decision support tool for RBD and PD. The motor symptoms associated with RBD are subtle, which makes it challenging to detect and monitor granular changes. Our analysis is aimed at testing an initial hypothesis that acoustic analysis of speech signals can be used to detect and quantify the symptoms associated with RBD and PD. This is relevant as the objective quantification of symptom severity

using voice can potentially help identify and prioritize participants for PSG, and facilitate intervention in the prodromal stage of PD. The novelty of our work lies in assessing the relationship between speech impairment and symptom severity in isolated RBD, with a focus on motor symptoms, cognition, daytime sleepiness, depression, and the overall state of health. To the best of our knowledge, this is the largest dataset of smartphone-based voice recordings collected from a deeply phenotyped RBD cohort.

The paper is organized as follows. Section II presents the study design and clinical data. Section III describes the methodology focusing on voice segmentation, feature extraction and selection, statistical mapping, and model validation. Section IV presents out-of-sample results for discriminating the three groups (Controls, RBD, and PD) and predicting PROMs and clinician-rated scores. Conclusions are presented in Section V, and limitations of this study and plans for future work are discussed in Section VI.

## II. DATA

Voice recordings and clinical data were collected from participants enrolled in the Oxford Discovery Cohort (further details are discussed in Barber et al. [32]; Baig et al. [33]; Lo et al. [13]). PWP met the United Kingdom PD Brain Bank criteria for probable PD [34]. We included PWP for whom the probability of PD was at least 90% (as ascertained by a trained researcher) at their most recent clinic visit. Participants with isolated RBD were included if their PSG provided evidence supportive of their clinical diagnosis, in keeping with the International Classification of Sleep Disorders criteria [35]. The study was prospectively approved by the local UK National Health Service Ethics research ethics committee (10/H0505/71 and 16/SC/0108), in adherence with national legislation and the Declaration of Helsinki. All participants provided written informed consent before any study-related procedures.

We used data from a cohort of 539 participants, comprising 92 Controls, 112 RBD, and 335 PWP. Participants were provided smartphones installed with a fully customized smartphone application that enabled the recording of a range of diverse modalities including voice, gait, balance, dexterity, reaction time, rest tremor, and postural tremor [12]. We focused only on the voice task in this study, for which the participants were provided with the instruction: "*Hold the phone to your ear, take a deep breath, and say 'aaah' at a comfortable and steady, tone and level, for as long as you can.*" The sustained vowel phonations "aaah" (International phonetic alphabet /a:/) were sampled at 44.1 kHz directly at the smartphone, and the recordings were encrypted, timestamped, and uploaded to a secure online database.

During their in-clinic visit, in conjunction with clinical assessments, participants performed the voice task under the supervision of a trained researcher. Moreover, participants were encouraged to take the smartphones home to perform the voice task up to four times a day, for seven days. The duration of each voice task was 20 seconds. Smartphone data collected during the first clinic visit and subsequent home recordings (performed within three months of their clinic visit) were used for analysis. Our findings are thus not dependent on the voice task being performed by participants under supervision in clinic. In total, we identified 4242 phonations ($n_{Controls}$ = 688, $n_{RBD}$ = 1359, $n_{PD}$ = 2195) from participants that fulfilled the above inclusion criteria.

Along with speech, we collected various established clinically validated metrics that are either expert rater-based or PROMs-based, including the MDS-UPDRS (we report both the motor MDS-UPDRS (part III, motor examination) and the total MDS-UPDRS), Montreal cognitive assessment, Epworth sleepiness scale, Beck depression inventory, and visual analogue scale (details for each outlined below). In all cases, the clinical assessment and the self-reports were collected in addition to the speech data. Basic demographics and participant information are summarized in Table I.

TABLE I
SUMMARY DEMOGRAPHICS AND PARTICIPANT INFORMATION

| | Controls ($n$=92) | RBD ($n$=112) | PD ($n$=335) |
|---|---|---|---|
| Age (years) | 68.5 ± 13.2 | 68.4 ± 12.1 | 69.5 ± 13.3 |
| Gender (male/female) | 73/19 | 97/15 | 206/129 |
| Total number of /a:/ phonations (male/female) | 688 (583/105) | 1359 (1154/205) | 2195 (1311/884) |
| Years since PD diagnosis and smartphone assessment | N/A | N/A | 3.93 ± 2.2 |
| Motor MDS-UPDRS | 2.0 ± 3.0 *($n$=54) | 5.0 ± 6.0 | 28.0 ± 17.0 |
| Total MDS-UPDRS I-III | 7.0 ± 7.0 *($n$=30) | 16.0 ± 12.0 | 49.0 ± 26.0 |
| MoCA | 27.0 ± 3.0 *($n$=50) | 26.0 ± 3.0 | 26.0 ± 5.0 |
| ESS | 5.0 ± 5.0 *($n$=51) | 6.0 ± 7.0 | 7.0 ± 6.8 |
| BDI | 2.0 ± 5.5 *($n$=51) | 8.0 ± 12.0 *($n$=109) | 8.0 ± 8.0 *($n$=327) |
| EQ-5D-3L VAS | 85.0 ± 10.0 *($n$=50) | 80.0 ± 20.0 *($n$=111) | 70.0 ± 20.0 *($n$=332) |
| RBDSQ | N/A | 10.0 ± 3.0 | N/A |

Summary statistics are presented in the form median ± interquartile range. Abbreviations used: RBD, rapid eye movement sleep behaviour disorder; PD, Parkinson's disease; MDS-UPDRS, Movement Disorder Society (MDS)-sponsored revision of the Unified Parkinson's disease rating scale; MoCA, Montreal cognitive assessment; ESS, Epworth sleepiness scale; BDI, Beck depression inventory; VAS, Visual analogue scale; RBDSQ, RBD screening questionnaire. We have included the number of participants ($n$) for the cases as we do not have entries for all participants in that group.

### A. MDS-UPDRS

The MDS-UPDRS is one of the most widely used measures to quantify the severity of PD [2]. In this study, we use the motor MDS-UPDRS (the third subscale of the MDS-UPDRS, which is also referred to as MDS-UPDRS part III) and the total MDS-UPDRS, which constitutes of the following four subscales: (I) nonmotor elements of PD, (II) nonmotor experiences of daily living, (III) motor examination, and (IV)

motor complications. The motor MDS-UPDRS is administered by a clinician and focuses on assessing the severity of motor symptoms. It comprises 33-items, whereby each item is scored using the following points scheme: normal (0), slight (1), mild (2), moderate (3), and severe (4). The maximum value of the motor MDS-UPDRS is 132 points, and a higher score represents more severe impairment. For the Discovery cohort, part (IV) of the MDS-UPDRS was administered only for the PD cohort. In this study, the total MDS-UPDRS was computed as the sum of the first three subscales (which we refer to as total MDS-UPDRS I-III).

### *B. MoCA*

The Montreal Cognitive Assessment (MoCA) is a brief 10-minute screening test, which exhibits high sensitivity and specificity in detecting the signs of Mild Cognitive Impairment (MCI), which is a clinical state that may evolve to dementia [36]. The MoCA is a 30-point test that evaluates: short-term memory recall, visuospatial abilities, multiple aspects of executive functions, attention, concentration, working memory, language, and orientation to time and place. A lower score is associated with a higher likelihood of MCI. We used the total MoCA score that was adjusted for education, whereby participants with $\leq$ 12 years of education were assigned an additional MoCA point [37].

### *C. ESS*

The Epworth Sleepiness Scale (ESS) is a PROMs-based questionnaire that assesses 'daytime sleepiness' [38]. The test comprises 8-items, each rated on a 4-point scale (with 0 denoting 'would never doze' and 3 denoting 'high chance of dozing'), and the total ESS has a range of 0–24.

### *D. BDI*

The Beck Depression Inventory (BDI) is a patient self-reported test that is used to measure the symptoms and severity of depression in persons aged $\geq$ 13 years. The BDI was introduced in 1961 and has since undergone multiple revisions [39]. In this study, we use BDI-II, which is a 21-item multiple-choice inventory, in which each item is rated out on a 4-point scale (0 to 3, where 3 indicates an extreme form of each symptom) [40]. The total BDI-II score has a range of 0 to 63, and the interpretation of this score is based on the following guidelines: minimal range (0–13), mild depression (14–19), moderate depression (20–28), and severe depression (29–63).

### *E. EQ-5D-3L VAS*

The Visual Analogue Scale (VAS) is a self-reported test used to measure the participants' health status on the day of the interview [41]. Participants were asked to mark their health status on a vertical scale, whereby the 'Worst imaginable health state' corresponds to a score of 0 and 'Best imaginable health state' equates to a score of 100.

### *F. RBDSQ*

The RBD Screening Questionnaire (RBDSQ) is a PROMs-based instrument which is based on a 10-item questionnaire (with each response being either 'yes' or 'no') [42]. The range is 0 to 13 points, where a higher score is associated with a higher likelihood of clinical RBD. RBDSQ assesses sleep behavior, focusing on a range of different nocturnal aspects, including frequency and content of dreams, nocturnal motor behavior, injuries, nocturnal awakenings, disturbed sleep, and presence of any neurological disorder. Using a cut-off of 5 points (as a positive diagnosis of RBD), a sensitivity of 96% and a specificity of 56% in discriminating RBD versus controls has been reported [42].

## III. METHODS

Our methodology is aimed at characterizing each sustained vowel phonation to extract informative acoustic measures (also referred to as *features*), determining a robust feature subset using feature selection algorithms, and mapping the selected feature subset onto the clinical outcomes of interest. A schematic diagram illustrating the different key stages of our modeling framework is provided in Fig. 1.

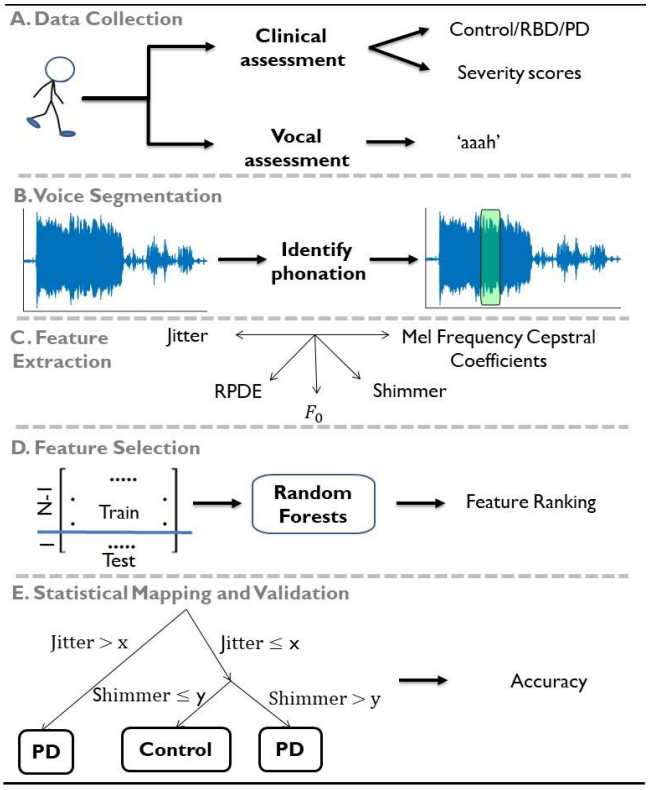


**FIGURE 1. Schematic diagram illustrating the acquisition of the clinical and voice data, and major steps involved in the analyses.**

Following the confirmation of study group (Control/RBD/PD), quantification of symptom severity (using the clinical scores) and vocal assessment, as shown in Fig. 1 (step A), the first step of our analyses undertook *voice segmentation*, which was aimed at identifying the voice segment that corresponds to the sustained vowel phonation

from the complete duration of the voice recording (step B in Fig. 1). Using the segmented phonation, we performed *feature extraction*, which was aimed at characterizing different acoustic measures of the signal (step C in Fig. 1). The feature matrix (and corresponding labels) were split into training data and testing data using a leave-one-subject-out cross-validation scheme, whereby all recordings except recordings from one participant were used for training the model and for identifying the most salient set of features, i.e., *feature selection* (step D in Fig. 1). The process was repeated, iteratively leaving the recordings from each participant out. We then performed *statistical mapping* to establish the relationship between the input features and the target label, whereby using the trained model, predictions were generated for the test dataset (one-by-one, for all participants) and the model accuracy was *validated* using a performance score (step E in Fig. 1). We now describe the different steps of our methodology in more detail below.

### A. Voice segmentation

Compared to supervised laboratory collected recordings, data acquired under non-controlled, free-living conditions yields findings that are more scalable to the real-world environment. Collecting data under non-controlled settings can, however, give rise to data quality issues, such as background noise, unexpected user behaviors, etc., which can potentially reduce the interpretability and reliability of the analysis. To tackle this issue, we developed an automated voice segmentation algorithm to identify the most stable single 2-second segment of sustained phonation from the voice recordings. The segmentation was based on the analysis of changes in fundamental frequency over different parts of the voice signal. The fundamental frequency (F0) of the speech signals was computed using the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) algorithm [43], which we had previously demonstrated to be the most accurate F0 estimation algorithm in sustained vowel /a:/ signals [44].

### B. Feature extraction

We characterized each sustained vowel /a:/ phonation using custom-built signal processing algorithms to compute 337 acoustic measures. We have developed a toolkit containing known and novel acoustic measures which we have refined over the years, specifically for processing sustained vowel /a:/ phonations [3], [16], [45], [46]. Briefly, these acoustic measures aimed to quantify the deviation from vocal fold periodicity (in terms of frequency the *jitter variants* and in terms of amplitude the *shimmer variants*), acoustic/turbulent noise, and articulator placement. For the physiological background, rationale, and detailed algorithmic expressions for the computation of the acoustic measures please refer to our previous studies [6], [13-15]. The MATLAB source code for the computation of the acoustic measures is freely available on the author's (AT) website: https://www.darth-group.com/software. Applying the speech signal processing algorithms to the study cohort gave rise to a 4242×337 feature matrix. These acoustic measures are summarized in Table II, whereas Table III for convenience summarizes the key acoustic aspects we aim to quantify using algorithmic processing and the corresponding acoustic measures. We remark there are different approaches to categorizing the acoustic measures, and the proposed approach serves as a useful methodological summary perspective. Also note that some acoustic measures to a certain extent, quantify aspects of more than one of the assigned categories.

TABLE II
BREAKDOWN OF THE 337 ACOUSTIC MEASURES USED IN THIS STUDY

| Family of acoustic measures | Brief description | Number of measures |
|---|---|---|
| Jitter variants | F0 perturbation | 21 |
| Shimmer variants | Amplitude perturbation | 22 |
| Harmonics to Noise Ratio (HNR) and Noise to Harmonics Ratio (NHR) | Signal to noise, and noise to signal ratios | 4 |
| Glottis Quotient (GQ) | Vocal fold cycle duration changes | 3 |
| Glottal to Noise Excitation (GNE) | Extent of noise in speech using energy and nonlinear energy concepts | 6 |
| Vocal Fold Excitation Ratio (VFER) | Extent of noise in speech using energy, nonlinear energy, and entropy concepts | 6 |
| Empirical Mode Decomposition Excitation Ratio (EMD-ER) | Signal to noise ratios using EMD-based energy, nonlinear energy and entropy | 6 |
| Mel Frequency Cepstral Coefficients (MFCC) | Amplitude and spectral fluctuations | 84 |
| Wavelet-based coefficients | Amplitude, scale, and envelope fluctuations quantified using wavelet coefficients | 182 |
| Pitch Period Entropy (PPE) | Inefficiency of F0 control | 1 |
| Detrended Fluctuation Analysis (DFA) | Stochastic self-similarity of turbulent noise | 1 |
| Recurrence Period Density Entropy (RPDE) | Uncertainty in estimation of fundamental frequency | 1 |

Algorithmic expressions for the 337 acoustic measures summarized here are described in detail in [3], [19], [45], [46], [53]. The MATLAB source code for the computation of the acoustic measures is freely available on the author's (AT) website: https://www.darth-group.com/software. F0 refers to fundamental frequency estimates, here computed using SWIPE [43].

### C. Feature exploration and statistical analysis

We explored the data using standard visualization tools in the form of violin plots to get a succinct representation of the underlying variable distributions. Subsequently, we computed correlation coefficients to express the statistical association between the acoustic measures and the clinical scales. We used the non-parametric Spearman correlation coefficient to account for a generic approach which does not require data normality, and computed statistical significance at the 95% level (p-values) for the null hypothesis that the acoustic measures were not statistically correlated with the clinical scales. We considered a relationship to be *statistically strong*

when the magnitude of the correlation coefficient R is at least 0.3, using the empirical rule of thumb in biomedical applications [47].

TABLE III
KEY ACOUSTIC ASPECTS AND CORRESPONDING ACOUSTIC MEASURES

| Key acoustic aspect quantified | Acoustic measures used |
|---|---|
| Deviations in retaining stable F0 and F0 variability | Jitter variants, Pitch Period Entropy (PPE), Recurrence Period Density Entropy (RPDE), Glottis Quotient (GQ), wavelet-based coefficients for F0 variability assessment |
| Deviations in retaining stable amplitude | Shimmer variants |
| Signal to noise ratio, quantifying excessive level of acoustic noise | Harmonics to Noise Ratio (HNR) and Noise to Harmonics Ratio (NHR), Detrended Fluctuation Analysis (DFA), Glottal to Noise Excitation (GNE), Vocal Fold Excitation Ratio (VFER), Empirical Mode Decomposition Excitation Ratio (EMD-ER) |
| Envelope (low frequency, general waveform aspect) | Lower MFCCs (and corresponding delta and delta-delta MFCCs) |
| High frequency/harmonic components | Higher MFCCs (and corresponding delta and delta-delta MFCCs) |

### D. Feature selection

A well-known problem in practical data analytics is the *curse of dimensionality*: a large number of features increases the noise in the dataset and may be detrimental in the statistical learning process [48]. Occam's razor dictates that we should aim to determine the most *parsimonious* statistical model, i.e. develop a statistical learning model that is maximally predictive with the minimum number of features. There are many different strategies to perform Feature Selection (FS); for an overview please refer to Guyon et al. [49]. Here, we used the importance scores from the Random Forests (RF) algorithm (see the following section) to rank the features and identify a robust subset. This embedded FS approach has the advantage that it is integral to the RF model building process alleviating the need for an additional external step towards FS, and has shown promising results in diverse applications [50].

### E. Statistical mapping

There are many statistical mapping algorithms in the literature, and this continues to be an active area of research. Here, we used RF [51] following the recommendation of Hastie et al. that tree-based ensembles are the *best off-the-shelf classifiers* [48]. A key competitive advantage of RF over some competing advanced statistical learning algorithms is that RF is very robust to the choice of hyperparameters (number of trees and number of features over which to optimize). We used the standard settings for these hyperparameters following Breiman's recommendations [51]: 500 trees and the square root of the number of features for split point selection at each node. Moreover, to tackle class imbalance, the votes cut-off for the classes was changed such that the minority class had a lower cut-off (directly proportional to the number of observations in that class) [52].

### F. Model performance and validation

To assess the statistical model performance and investigate its performance in unseen data, we used the standard Leave-One-Subject-Out (LOSO) Cross-Validation (CV) approach. Specifically, the statistical learning model was trained using the samples of the $N$-1 unique participants, and tested on the performance of correctly estimating the data for the participant that was not used in the training phase. Using LOSO, the process of model training and predicting was repeated for a total of $N$ times (one time for each participant). Using all $N$ labels and corresponding predictions, we report the sensitivity and specificity for pairwise discriminations, and the Mean Absolute Error (MAE) when referring to estimating the clinical scores. The MAE values are summarized in the form median ± Interquartile Range (IQR).

To discriminate the three groups (Controls, RBD and PD), using only the features extracted from sustained phonations, we performed the following pairwise comparisons: (1) Controls versus PD, (2) Controls versus RBD, and (3) RBD versus PD. For each pairwise comparison, we employed an ensemble of classification trees using: (1) all available recordings, (2) only male recordings, and (3) only female recordings. Moreover, to investigate the effect of the data size on classification accuracy, we performed the analyses using: (i) only a single voice recording per participant (first voice recording collected from each participant), and (ii) total number of recordings contributed by a given participant. Since different participants contributed a different number of recordings in the testing scenario (ii), we performed model validation such that each participant was assigned equal weight during the model validation. Specifically, for a given participant, we used a majority voting scheme to determine if the majority of recordings were classified as belonging to either class 1 or class 2, assigning the final estimate to the majority class for that participant. This resulted in one label and one classification per participant, which was subsequently used for assessing the model performance. Additional details pertaining to the analyses can be found in [3], [19], [45], [46], [53], and references therein.

## IV. RESULTS

Participants from the three groups (Controls, RBD, and PD) were age-matched. Pairwise comparisons of age distributions (Controls vs PD, Controls vs RBD, and RBD vs PD) rejected the null hypothesis that the age distributions were significantly different (using a two-sided Kolmogorov-Smirnov test with 5% significance level). This helps garner confidence that the findings of our study are not biased due to the presence of presbyphonia as a potential confounding factor. The Controls and RBD groups were male dominant. We start our exploration by visualizing the underlying distributions of the clinical scales for the three cohorts (see Fig. 2).

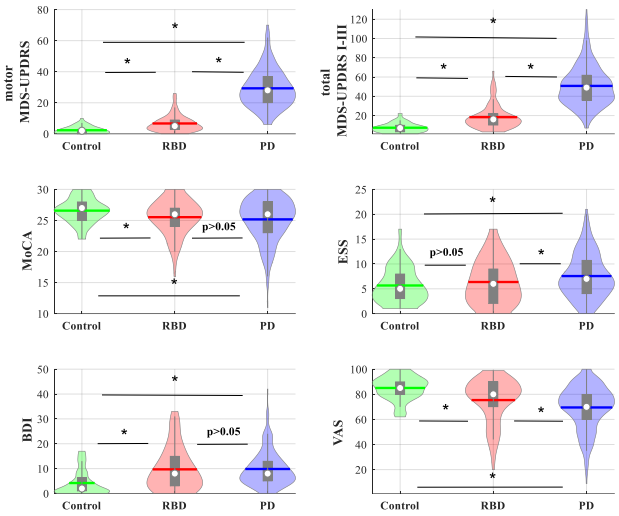


**FIGURE 2. Violin plots summarizing the distributions of the key clinical metrics and comparing the three groups. The boxplot is embedded within each violin plot, where the white circle denotes the median and the grey box denotes the 25th percentile (lower end) and 75th percentile (upper end). The horizontal line within each violin plot denotes the mean.**
Clinical metrics are analysed across the three groups (Controls vs RBD, RBD vs PD, and Controls vs PD) using the Mann–Whitney U test. Statistically significant findings ($p < 0.05$) are marked using *.

The difference in all clinical metrics for the control and PD cohorts were found to be statistically significant, while this was not the case for MoCA and BDI for the RBD and PD cohorts. To account for potential group differences in sex, we stratified the data to present the results separately for each pairwise group comparison, using all recordings, only female recordings, and only male recordings. We next focus on investigating the pairwise discrimination of the three cohorts using speech signals. Using only a single recording per participant ($n_{Controls}$ = 92, $n_{RBD}$ = 112, $n_{PD}$ = 335), the out-of-sample classification accuracy was slightly higher for RBD versus PD, compared to the accuracy obtained in discriminating Controls vs PD, and Controls versus RBD, as shown in Table IV.

TABLE IV
DISCRIMINATION ACCURACIES FOR THE 3 PAIRWISE COMPARISONS USING THE LEAVE-ONE-RECORDING-OUT CROSS-VALIDATION SCHEME (USING ONLY 1 RECORDING PER PARTICIPANT)

| **Discrimination accuracy** | | **Sensitivity (%)** | **Specificity (%)** |
|---|---|---|---|
| Controls vs PD | All | 62.1% | 56.5% |
| | Male | 61.7% | 58.9% |
| | Female | 47.3% | 36.8% |
| Controls vs RBD | All | 56.3% | 70.7% |
| | Male | 57.7% | 69.9% |
| | Female | 40.0% | 42.1% |
| RBD vs PD | All | 66.9% | 66.1% |
| | Male | 60.7% | 70.1% |
| | Female | 59.7% | 66.7% |
| **No. of recordings** | $n_{All}$ | $n_{Male}$ | $n_{Female}$ |
| Controls | 92 | 73 | 19 |
| RBD | 112 | 97 | 15 |
| PD | 335 | 206 | 129 |

We chose the recording corresponding to the first speech test performed by each participant in Table IV. For all three pairwise comparisons, the accuracy obtained using all recordings and only male recordings were rather similar, while the accuracy using only female recordings were poor. This can be attributed to the fact that both RBD and Control cohorts comprised very few female participants, and thus the analyses using only female recordings are likely to be less reliable.

The out-of-sample classification accuracy obtained using the total number of available recordings and a majority assignment scheme (to assign equal weight to each participant during the model validation) is presented in Table V. While we were able to distinguish RBD participants from controls and PD with decent accuracies (Table V), the discrimination accuracy for Controls vs PD, was surprisingly poor. Although this requires further investigation, a potential reason for poor discrimination accuracy using the control recordings could be that compared to the other two cohorts, the number of control recordings were about half and one-third of the total number of recordings from RBD and PD participants, respectively. Moreover, only 39 controls contributed more than one speech recording, as opposed to 76 RBD and 126 PD participants who performed multiple speech tests.

In terms of discriminating Controls vs RBD, the results of this study (as presented in Table V, sensitivity 56.3% and specificity 70.7%) are in broad agreement with previous findings that were based on a smaller cohort which had reported sensitivity 69.8% and specificity 64.7% [30].

To further explore reasons for the poor discrimination accuracy, using all recordings for Controls vs PD, we undertook additional analyses employing following schemes to alleviate class imbalance issues: class weights, undersampling the majority class, and RUSBoost [52], [54]. However, the discrimination accuracy was not noticeably better using these schemes.

TABLE V
DISCRIMINATION ACCURACIES FOR THE 3 PAIRWISE COMPARISONS USING THE LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION SCHEME (USING TOTAL NUMBER OF RECORDINGS AND MAJORITY ASSIGNMENT)

| **Discrimination accuracy** | | **Sensitivity (%)** | **Specificity (%)** |
|---|---|---|---|
| Controls vs PD | All | 59.4% | 67.4% |
| | Male | 55.3% | 72.6% |
| | Female | 48.1% | 36.8% |
| Controls vs RBD | All | 60.7% | 69.6% |
| | Male | 59.8% | 74.0% |
| | Female | 46.7% | 26.3% |
| RBD vs PD | All | 74.9% | 73.2% |
| | Male | 74.8% | 75.3% |
| | Female | 51.9% | 46.7% |
| **No. of recordings** | $n_{All}$ | $n_{Male}$ | $n_{Female}$ |
| Controls | 688 | 583 | 105 |
| RBD | 1359 | 1154 | 205 |
| PD | 2195 | 1311 | 884 |

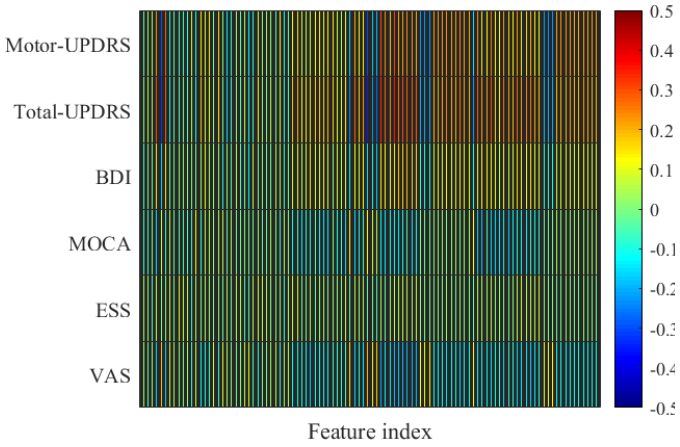


**FIGURE 3. Heatmap summarizing the statistical associations for all participants across the 337 features with the 6 clinical scales.**

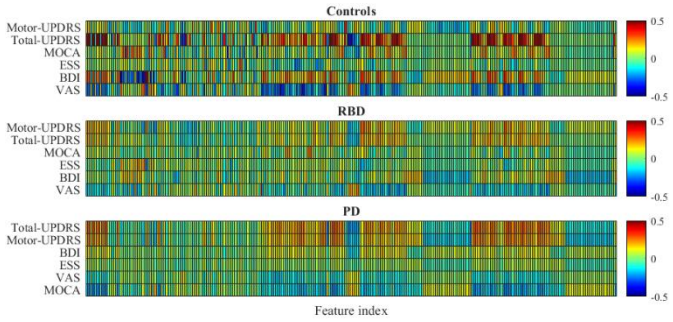


**FIGURE 4. Heatmap summarizing the statistical association for the stratified group cohorts across the 337 features with the 6 clinical scales.**

Subsequently, we investigated the statistical associations of the features with the clinical scales; to keep those concise the results are summarized in Fig. 3 for all participants, and then in Fig. 4 stratifying the data for the three cohorts. Collectively, the findings in Figs. 3 and 4 suggest there are some statistically strong associations between the acoustic measures and the clinical scales. In Figs. 3 and 4, the magnitude of the Spearman correlation coefficients (using only one speech recording per participant) was less than 0.5 and hence we have compressed the scale presented in the range [-0.5 0.5], whereby the order of the 337 features in the heatmap follows the presentation in Table II. In a few cases, strong correlations were revealed only after stratifying the original dataset into the group cohorts, which motivates the need to develop stratified cohort-based models to estimate the different clinical scales using speech. We defer more detailed elaboration on the most strongly associated features with the clinical scales and cross-comparisons for the Discussion.

Table VI presents the out-of-sample LOSO results for each of the clinical scales for the three cohorts and also for the three groups collectively. We have followed the methodology outlined above assessing the performance of the classifier both when using a single recording per participant, and also all available recordings with a majority voting scheme on LOSO validation. The results were similar, and here we present only the findings with a single recording per participant.

TABLE VI
OUT OF SAMPLE MAE PERFORMANCE ACROSS CLINICAL SCALES

| Clinical scale | Cohort | | |
|---|---|---|---|
| | Controls ($n$=92) | RBD ($n$=112) | PD ($n$=335) |
| Motor MDS-UPDRS | 1.0 ± 2.0 | 2.5 ± 5.0 | 8.0 ± 9.0 |
| Total MDS-UPDRS I-III | 1.0 ± 3.0 | 6.0 ± 8.0 | 14.0 ± 18.0 |
| MoCA | 1.0 ± 2.0 | 2.0 ± 2.0 | 2.0 ± 3.0 |
| ESS | 2.0 ± 2.0 | 3.0 ± 4.0 | 3.0 ± 4.0 |
| BDI | 1.0 ± 4.0 | 5.0 ± 8.0 | 4.0 ± 6.0 |
| VAS | 6.5 ± 5.0 | 10.0 ± 18.0 | 10.0 ± 15.0 |

The out of sample Mean Absolute Error (MAE) performance reported here was computed using the leave-one-subject-out cross-validation scheme: this corresponds to the MAE between the ground truth and the estimates for each individual in the cohort (e.g. we computed 335 entries for PD). Subsequently we need to succinctly summarize these MAE entries and here we report findings in the form median ± IQR. These results were determined by feeding into RF all 337 features and also exploring whether feeding in progressively the top 1…25 features resulting from the RF importance scores for each sub-problem (indicatively, we illustrate these for motor MDS-UPDRS in Fig. 5).

Indicatively, we illustrate in Fig. 5 performance of RF in predicting the motor MDS-UPDRS. The out-of-sample model performance, as quantified using the MAE, is shown as a function of the number of most salient features used during the modelling. The order of the presented features was determined using the ranked RF importance scores. For brevity, we only illustrate the results for motor MDS-UPDRS.

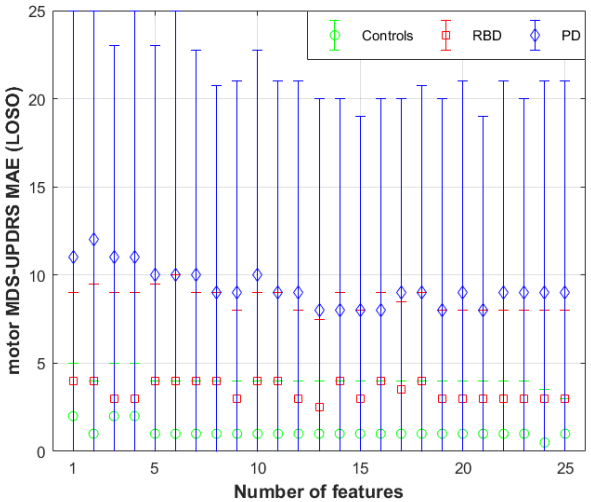


**FIGURE 5. LOSO MAE performance of the RF in predicting the motor MDS-UPDRS as a function of the number of features presented into the classifier. The corresponding symbol in each case indicates the median and the bars the IQR.**

Finally, for the RBD cohort, we explored how the acoustic measures relate to the RBDSQ score. Table VII presents the correlation coefficients of the ten most strongly associated acoustic measures with the RBDSQ score. Similarly, to the preceding analyses, we have aimed to estimate RBDSQ presenting the acoustic measures into RF for evaluating the

model performance in a LOSO framework. Encouragingly, we have found that the RBDSQ can be estimated accurately for the RBD cohort ($n_{RBD}$ = 112) with a LOSO MAE of (median ± IQR) 1 ± 1 RBDSQ points.

TABLE VII
CORRELATION COEFFICIENTS OF ACOUSTIC MEASURES WITH RBDSQ

| Acoustic measure | Correlation coefficient |
|---|---|
| Standard deviation of the 0th delta-delta MFCC | 0.260 |
| $VFER_{mean}$ | -0.253 |
| Average of the 9th delta MFCC | 0.248 |
| $VFER_{entropy}$ | -0.244 |
| Standard deviation of the 3rd MFCC | 0.241 |
| Standard deviation of the 0th MFCC | 0.220 |
| $EMD\text{-}ER_{SNR,SEO}$ | -0.217 |
| Standard deviation of the 0th delta MFCC | 0.215 |
| Standard deviation of the 3rd delta-delta MFCC | 0.206 |
| Average of the 7th delta MFCC | -0.196 |

We present only the 10 most strongly associated acoustic measures for brevity. In all cases the correlations were statistically significant ($p<0.05$).

## V. CONCLUSIONS

This study aimed to provide the first comprehensive investigation of a diverse range of PD and RBD clinical scales when using smartphone-based speech signal analysis. We have found that speech can be used to estimate diverse PD and RBD clinical scales with reported MAE that would make these estimations clinically meaningful. We demonstrated that RBDSQ can be estimated very accurately with a MAE of (median ± IQR) 1 ± 1 points. Given that RBD is a group that may convert to PD and that the RBDSQ quantifies RBD symptoms, this finding may have important implications towards early assessment of prodromal symptoms in PD prior to clinical diagnosis. Moreover, we were able to distinguish RBD participants from both controls (sensitivity 60.7%, specificity 69.6%) and PD (sensitivity 74.9%, specificity 73.2%). These results could potentially indicate that the vocal deficits in participants with isolated RBD might be different than those with PD. These findings warrant longitudinal studies to investigate speech impairment in participants with RBD. While previous studies have typically focused on speech analysis for PD, this study demonstrates that speech provides the means towards clinically meaningful insights into symptom severity displayed across the spectrum of both PD and RBD. These results from a deeply clinically phenotyped cohort highlights that speech can potentially be used as a digital biomarker for prodromal PD.

We emphasize that the PD cohort were at the early stages of the disease with relatively mild symptoms as summarized in MDS-UPDRS (see Table I). Moreover, whilst previous speech-RBD studies have employed lab-quality recordings, we felt it was imperative to use recordings collected under realistic environment settings to address issues regarding scalability and generalizability of previous findings. It is for this reason that we collected voice recordings under clinic- and home-based settings via smartphones, from one of the largest cohorts of RBD and PD participants. The data were collected by participants themselves using a wide variety of off-the-shelf consumer-grade smartphones (manufactured by major international brands).

We explored the statistical associations (using Spearman correlation coefficients) of 337 features, which have been used in similar problems when processing sustained vowel /a:/ phonations in PD, with six widely used PD clinical scales (see Fig. 3). We confirmed some of our previous findings [3], [16], finding statistically strong associations ($|R|>0.3$) between some of the acoustic measures and the MDS-UPDRS (both motor MDS-UPDRS and total MDS-UPDRS I-III). Interestingly, for some of the clinical scales, we observed that statistical correlation became more pronounced in stratified groups (see Fig. 4). Further work is needed to verify these findings in larger Control, RBD and PD cohorts.

RF derived feature rankings were sub-problem specific and did not generalize across problems (results not shown), verifying what could have been expected also when visualizing the statistical correlations summarized in Fig. 4. This tacitly suggests there are different underlying properties quantified by the acoustic measures which were best tailored for the estimation of the different clinical scales. Overall, we found that a proportion of features from the VFER-family, MFCCs and wavelet-based acoustic measures were statistically strongly correlated with the clinical scales (Fig. 4) and were highly ranked using the RF importance scores. This is broadly in agreement with our previous findings in related PD applications [3], [19], [21].

## VI. LIMITATIONS AND FUTURE WORK

Despite the promising findings reported herein, there are some limitations of this study. Firstly, the quality of voice samples collected using smartphones under clinic- and home-based settings is likely to be of relatively worse quality compared to data collected under acoustically highly controlled lab-settings (e.g., double-walled sound booths), which potentially translates into lower discriminatory accuracy for the cohorts investigated. Secondly, this study relies on acoustic signal analysis using only one type of sustained phonation (“aaah”), which may not adequately encapsulate the whole spectrum of speech symptoms in RBD and PD. It is plausible that acoustic analysis based on a multitude of sustained phonation types, syllable repetition, and monologue, may improve the efficacy of the biomarker and provide a more complete understanding of the degree of speech impairment in PD, such as soft speech (hypophonia), monotonous speech with the lack of inflection (aprosody), and dysarthria in the form of inability to separate syllables clearly (tachyphemia). Thirdly, we collected data from only three groups (controls, RBD, and PD), thereby not accounting for other parkinsonism and tremor disorders that may also exhibit comparable patterns of impairment in speech. Therefore, the

extent of our claims on the basis of the available data is restricted to the differentiation of the three cohorts. The development of a robust, reliable clinical decision support tool towards differential diagnosis would require the use of a large sample size across a wider range of related neurodegenerative disorders. Finally, although we validated our statistical framework on an independent test dataset, we used data from only one cohort (Discovery cohort). Validation based on an external independent cohort would have provided additional reliability to these findings. Future studies could address some of the aforementioned limitations. An interesting line of future work would be to longitudinally monitor speech, along with other motor and non-motor symptoms, with a particular focus on participants with RBD who eventually convert to an overt neurodegenerative disease. We envisage the findings of this work would contribute towards the risk stratification of individuals who are at the risk of developing PD and assist in remote longitudinal monitoring of PD symptoms. Overall, this study extends the increasing evidence presented in the research literature capitalizing on biomedical speech signal processing towards the objective assessment of RBD and PD.

## ACKNOWLEDGEMENTS

We are grateful to the participants of this study for their commitment and time, that made this research possible. We extend our sincere thanks to members of the Oxford Discovery team who assisted in data acquisition during research clinics, and Max A. Little and Andong Zhan for their collaboration with the smartphone app. We would like to thank the three anonymous Reviewers for their comments, which helped improve the quality of this manuscript.

## REFERENCES

[1] E. Ray Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurol.*, vol. 17, no. 11, pp. 939–953, 2018, doi: 10.1016/S1474-4422(18)30295-3.

[2] C. G. Goetz *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.*, vol. 23, no. 15, pp. 2129–2170, 2008, doi: 10.1002/mds.22340.

[3] A. Tsanas, "Accurate telemonitoring of Parkinson's disease using nonlinear speech signal processing and statistical machine learning," University of Oxford, 2012.

[4] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998, doi: 10.1155/1999/327643.

[5] Z. Pirtošek, O. Bajenaru, N. Kovács, I. Milanov, M. Relja, and M. Skorvanek, "Update on the Management of Parkinson's Disease for General Neurologists," *Parkinsons. Dis.*, vol. 2020, 2020, doi: 10.1155/2020/9131474.

[6] B. Post, M. P. Merkus, R. M. a de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?," *Mov. Disord.*, vol. 20, no. 12, pp. 1577–84, Dec. 2005, doi: 10.1002/mds.20640.

[7] A. Tsanas *et al.*, "Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder," *J. Affect. Disord.*, vol. 205, pp. 225–233, 2016, doi: 10.1016/j.jad.2016.06.065.

[8] M. Faurholt-Jepsen, S. Brage, M. Vinberg, and L. V Kessing, "State-related differences in the level of psychomotor activity in patients with bipolar disorder - Continuous heart rate and movement monitoring," *Psychiatry Res.*, vol. 237, pp. 166–174, 2016, doi: 10.1016/j.psychres.2016.01.047.

[9] N. Palmius *et al.*, "A multi-sensor monitoring system for objective mental health management in resource constrained environments," in *IET Appropriate healthcare technologies for low resource settings*, 2014.

[10] A. J. Espray *et al.*, "Technology in Parkinson disease: Challenges and Opportunities," *Mov. Disord.*, vol. 31, no. 9, pp. 1272–1282, 2017, doi: 10.1002/mds.26642.

[11] N. Mahadevan *et al.*, "Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device," *npj Digit. Med.*, vol. 3, no. 1, 2020, doi: 10.1038/s41746-019-0217-7.

[12] S. Arora *et al.*, "Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD," *Neurology*, vol. 91, no. 16, pp. E1528–E1538, 2018, doi: 10.1212/WNL.0000000000006366.

[13] C. Lo *et al.*, "Predicting motor, cognitive & functional impairment in Parkinson's," *Ann. Clin. Transl. Neurol.*, vol. 6, no. 8, pp. 1498–1509, 2019, doi: 10.1002/acn3.50853.

[14] A. Zhan *et al.*, "Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score," *JAMA Neurol.*, vol. 75, no. 7, pp. 876–880, 2018, doi: 10.1001/jamaneurol.2018.0809.

[15] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012, doi: 10.1109/TBME.2012.2183367.

[16] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity.," *J. R. Soc. Interface*, vol. 8, no. 59, pp. 842–855, 2011, doi: 10.1098/rsif.2010.0456.

[17] A. Tsanas, M. A. Little, and L. O. Ramig, "Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network," *IEEE Access*, vol. 9, pp. 11024–11036, 2021, doi: 10.1109/ACCESS.2021.3050524.

[18] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Robust parsimonious selection of dysphonia measures for telemonitoring of parkinson's disease symptom severity," in *Models and Analysis of Vocal Emissions for Biomedical Applications - 7th International Workshop, MAVEBA 2011*, 2011, pp. 169–172.

[19] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, 2014, doi: 10.1109/TNSRE.2013.2293575.

[20] S. Arora *et al.*, "Investigating Voice as a Biomarker for leucine-rich repeat kinase 2-Associated Parkinson's Disease," *J. Parkinsons. Dis.*, vol. 8, no. 4, pp. 503–510, 2018, doi: 10.3233/JPD-181389.

[21] S. Arora, L. Baghai-Ravary, and A. Tsanas, "Developing a large scale population screening tool for the assessment of Parkinson ' s disease using telephone-quality voice," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 2871–2884, 2019.

[22] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly, and P. J. Snyder, "Acoustic characteristics of Parkinsonian speech: A potential biomarker of early disease progression and treatment," *J. Neurolinguistics*, vol. 17, no. 6, pp. 439–453, 2004, doi: 10.1016/j.jneuroling.2004.06.001.

[23] A. Iranzo, J. Santamaria, and E. Tolosa, "Idiopathic rapid eye movement sleep behaviour disorder: Diagnosis, management, and the need for neuroprotective interventions," *Lancet Neurol.*, vol. 15, no. 4, pp. 405–419, 2016, doi: 10.1016/S1474-4422(16)00057-0.

[24] A. Iranzo *et al.*, "Neurodegenerative disorder risk in idiopathic REM sleep behavior disorder: Study in 174 patients," *PLoS One*, vol. 9, no. 2, 2014, doi: 10.1371/journal.pone.0089741.
[25] C. H. Schenck, B. F. Boeve, and M. W. Mahowald, "Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: A 16-year update on a previously reported series," *Sleep Med.*, vol. 14, no. 8, pp. 744–748, 2013, doi: 10.1016/j.sleep.2012.10.009.
[26] C. H. Schenck *et al.*, "Rapid eye movement sleep behavior disorder: Devising controlled active treatment studies for symptomatic and neuroprotective therapy-a consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group," *Sleep Med.*, vol. 14, no. 8, pp. 795–806, 2013, doi: 10.1016/j.sleep.2013.02.016.
[27] B. F. Boeve *et al.*, "Pathophysiology of REM sleep behaviour disorder and relevance to neurodegenerative disease," *Brain*, vol. 130, no. 11, pp. 2770–2788, 2007, doi: 10.1093/brain/awm056.
[28] A. Konka, J. Weedon, and N. A. Goldstein, "Cost-benefit analysis of polysomnography versus clinical assessment score-15 (CAS-15) for treatment of pediatric sleep-disordered breathing," *Otolaryngol. - Head Neck Surg. (United States)*, vol. 151, no. 3, pp. 484–488, 2014, doi: 10.1177/0194599814536844.
[29] J. Rusz *et al.*, "Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder," *Sleep Med.*, vol. 19, pp. 141–147, 2016, doi: 10.1016/j.sleep.2015.07.030.
[30] J. Rusz *et al.*, "Smartphone Allows Capture of Speech Abnormalities Associated with High Risk of Developing Parkinson's Disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1495–1507, 2018, doi: 10.1109/TNSRE.2018.2851787.
[31] A. Benba, A. Jilbab, S. Sandabad, and A. Hammouch, "Voice signal processing for detecting possible early signs of Parkinson's disease in patients with rapid eye movement sleep behavior disorder," *Int. J. Speech Technol.*, vol. 22, no. 1, pp. 121–129, 2019, doi: 10.1007/s10772-018-09588-0.
[32] T. R. Barber *et al.*, "Prodromal Parkinsonism and Neurodegenerative Risk Stratification in REM Sleep Behavior Disorder," *Sleep*, vol. 40, no. 8, pp. 11–13, 2017, doi: 10.1093/sleep/zsx071.
[33] F. Baig *et al.*, "Delineating nonmotor symptoms in early Parkinson's disease and first-degree relatives," *Mov. Disord.*, vol. 30, no. 13, pp. 1759–1766, 2015, doi: 10.1002/mds.26281.
[34] A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees, "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases," *J. Neurol. Neurosurg. Psychiatry*, vol. 55, no. 3, pp. 181–184, 1992, doi: 10.1136/jnnp.55.3.181.
[35] M. J. Sateia, "International classification of sleep disorders-third edition highlights and modifications," *Chest*, vol. 146, no. 5, pp. 1387–1394, 2014, doi: 10.1378/chest.14-0970.
[36] Z. S. Nasreddine *et al.*, "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment," *J. Am. Geriatr. Soc.*, vol. 53, no. 4, pp. 695–699, 2005, doi: 10.1111/j.1532-5415.2005.53221.x.
[37] G. Gagnon *et al.*, "Correcting the MoCA for education: Effect on sensitivity," *Can. J. Neurol. Sci.*, vol. 40, no. 5, pp. 678–683, 2013, doi: 10.1017/S0317167100014918.
[38] M. W. Johns, "A new method for measuring daytime sleepiness: The Epworth sleepiness scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991, doi: 10.1093/sleep/14.6.540.
[39] A. T. Beck, R. A. Steer, and M. G. Garbin, "Psychometric properties of the Beck Depression Inventory: twenty five years of evaluation," *Clin. Psychol. Rev.*, vol. 8, pp. 77–100, 1988, doi: 10.1016/j.psychres.2007.11.018.
[40] K. L. Smarr and A. L. Keefer, "Measures of depression and depressive symptoms," *Arthritis Care Res.*, vol. 63, pp. 454–466, 2011, doi: 10.1002/acr.20556.
[41] R. Brooks, "EuroQol: the current state of play," *Health Policy (New. York).*, vol. 37, no. 1, pp. 53–72, Jul. 1996, doi: 10.1016/0168-8510(96)00822-6.
[42] K. Stiasny-Kolster, G. Mayer, S. Schäfer, J. C. Möller, M. Heinzel-Gutenbrunner, and W. H. Oertel, "The REM sleep behavior disorder screening questionnaire - A new diagnostic instrument," *Mov. Disord.*, vol. 22, no. 16, pp. 2386–2393, 2007, doi: 10.1002/mds.21740.
[43] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music.," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–52, Sep. 2008, doi: 10.1121/1.2951592.
[44] A. Tsanas, M. Zañartu, M. a. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–2901, 2014, doi: 10.1121/1.4870484.
[45] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity," in *International symposium on nonlinear theory and its applications (NOLTA)*, 2010, no. September, pp. 457–460.
[46] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2013, pp. 37–40.
[47] A. Tsanas, M. A. Little, and P. E. Mcsharry, "A methodology for the analysis of medical data," in *Handbook of Systems and Complexity in Health*, J. P. Sturmberg and C. M. Martin, Eds. Springer, 2013, pp. 113–125.
[48] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of statistical learning*, 2nd editio. 2009.
[49] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
[50] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *J. Mach. Learn. Res.*, vol. 10, pp. 1341–1366, 2009.
[51] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1201/9780367816377-11.
[52] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," Berkeley, CA, 2004.
[53] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010, doi: 10.1109/TBME.2009.2036000.
[54] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

**SIDDHARTH ARORA** received a DPhil (PhD) from the University of Oxford (UK, 2013). He has since continued working at the University of Oxford, where he is currently a Departmental Lecturer. His research interests include biomedical signal processing, time series forecasting, chaos synchronization and statistical machine learning. He received the Martin Black Prize (2019) awarded by the Institute of Physics and Engineering in Medicine, most acclaimed lecturer award (2017) by the Oxford University Student Union, EPSRC Statistics and Machine Learning award (2015), and Lord Jenkins award (2010) by Somerville college at Oxford for academic performance. He sits on the Editorial Board of the journal Digital Biomarkers.

CHRISTINE LO is a Clinical Research Fellow at the Oxford Parkinson's Disease Centre in the University of Oxford. She has been working to identify novel clinical applications of smartphone data, contributed by participants in the Oxford Discovery study, utilising machine learning techniques. In keeping with her clinical interest in movement disorders, her research interests cover wearable devices and disease stratification in Parkinson's and related disorders.

**MICHELE HU** is Professor at the Nuffield Department of Clinical Neurosciences, University of Oxford and Honorary Consultant Neurologist at Oxford University Hospitals. She obtained her medical degree from King's College London, completing a PhD in Neuroscience at London University, and her neurology training at the Royal Free, National Hospital London, and Oxford University Hospitals.

Since her appointment as Movement Disorders Neurology Consultant in 2005, she has the advantage of having gained 15 years pragmatic experience looking after Parkinson's and other movement disorders patients, providing a unique insight that can be utilised to full advantage in the delivery of future treatments. She leads the clinical research program on the Oxford Parkinson's Disease Centre Discovery cohort (OPDC; www.opdc.ox.ac.uk), one of the largest longitudinal cohorts of Parkinson's patients in the world. Her group facilitates translational research in the field of longitudinal cohort studies and biomarkers for early and prodromal Parkinson's disease, with particular focus on REM sleep behaviour disorder (RBD), and how sleep affects neurodegeneration. Other key interests include the delivery of tractable, low cost, digital technology that has a real impact on patient's daily lives, and imaging the human brain from prodromal to established Parkinson's.

**ATHANASIOS TSANAS** (SM'19) received a BSc in Biomedical Engineering from the Technological Educational Institute of Athens (Greece, 2005), a BEng in Electrical Engineering and Electronics from the University of Liverpool (UK, 2007), an MSc in Signal Processing and Communications from the Newcastle University (UK, 2008), and a DPhil (PhD) in Applied Mathematics from the University of Oxford (UK, 2012). He continued working at the University of Oxford as a Research Fellow in Biomedical Engineering and Applied Mathematics (2012-2016), Stipendiary Lecturer in Engineering Science (2014-2016), and Lecturer in Statistical Research Methods (2016-2019). He joined the Usher Institute, Edinburgh Medical School, University of Edinburgh in January 2017 where he is a tenured Assoc. Prof. in Data Science and the Co-Director of TELESCOT. He leads the development and delivery of the 'Clinical Decision Support and Actionable Data Analytics' theme in the NHS Digital Academy, an innovative leadership programme jointly delivered with Imperial College London, which aims to train NHS leaders in the UK. His research interests are broadly in biomedical signal processing, time-series analysis, and statistical machine learning.

Dr. Tsanas received the Andrew Goudie award (top PhD student across all disciplines, St. Cross College, University of Oxford, 2011), the EPSRC Doctoral Prize award (2012) as one of only 8 Oxford PhD students across 11 departments, the young scientist award (MAVEBA, 2013), the EPSRC Statistics and Machine Learning award (2015), and won a 'Best reviewer award' from the IEEE Journal of Biomedical Health Informatics (2015). He was a key member of the Oxford Biomedical Engineering team that won the annual 2012 Physionet competition on 'Predicting mortality of ICU patients'. His research work has been highlighted in Renewable Energy and Global Innovations and in the media including Reuters. He sits on the Editorial Boards of JMIR Mental Health and JMIR mHealth and uHealth, and has served as Guest Editor for two special issues in other journals. He is a Senior Member of IEEE, a Fellow of the Higher Education Academy, and a Fellow of the Royal Society of Medicine.