Biomarkers Beyond the Clinic: Monitoring Parkinson's Disease with Smartphone Speech Samples

Stefan Bostock, Maarten De Vos, Detlef Wolf, Florian Lipsmeier, Michael Lindemann

Abstract — Diagnosis and monitoring of Parkinson's Disease is a costly, inconvenient and subjective process. The financial and temporal burden on patients, clinicians and pharmaceutical companies could be alleviated with accurate digital biomarkers obtained from widely available commercial devices. We show the delineation of male patients from healthy controls at an accuracy of 78.1% using sustained phonation recordings; made on the subjects smartphone outside of the clinic. We propose a methodology for optimizing and standardizing feature extraction from such speech samples, and utilize random forests to score patients on the Unified Parkinson's Disease Rating Scale. Further nuances of pathological speech analysis are also investigated such as the effects of analysis window size and location, as well as that of background noise.

***Index Terms*—Parkinson's disease, phonation, speech, random forests, telemedicine, eHealth, signal processing**

I. Introduction

Parkinson's disease (PD) is a neurological condition that predominantly affects older individuals. The disease is chiefly characterized by the loss of dopaminergic neurons in specific parts of the brain; outwardly portraying itself as: akinesia (slowed movement), muscle rigidity, tremor and often cognitive impairment. With the moderately high prevalence of approximately 1-2% over the age of 65, however, despite numerous studies, prevalence and how it differs by gender, age, nationality and ethnicity is disputed [1]–[4]. This is in part due to the fact that there is currently no definitive biomarker or imaging test for diagnosis. Clinical assessment is currently required, and often a patients response to dopamine agonists such as L-Dopa is the most informative diagnostic tool [5]. The situation is further confounded by the fact that other syndromes such as Dementia with Lewy Bodies (DLB) also exhibit 'Parkinsonism' [5].

The primary scoring tool for the disease is the Unified Parkinson's Disease Rating Scale (UPDRS), an assessment conducted by physicians that is comprised of four parts: mentation, behavior, and mood; activities of daily living; motor examination; and complications. Scores range between 0 and 176 with higher values for worse symptoms [6]. The scale has been shown to be a useful prognostic tool, however clinimetric studies dispute the objectivity and consistency of the system, with intra- and inter-rater variability [7]–[10]. Some reports suggest that this is particularly acute when assessing aphasia (speech disruption) and facial expressions [8], [9].

Patient speech in PD is of increasing interest to researchers. Although only contributing four potential points to the UPDRS score, it has obvious effects on patient quality of life, with around 30% calling it their most debilitating deficit, and it occurs in 70% to 90% of cases [11], [12]. Additionally, it has been hypothesized that due to degradation in cognitive capacity and stiffening/reduced control of the vocal folds, speech analysis may offer a sensitive, non-invasive staging and monitoring tool for PD.

Within the field of PD speech analysis, sustained phonation is one of the most thoroughly investigated areas. Asking the patient to annunciate a specified phoneme for an extended period of time can elucidate the stability and strength of signal as well as more profound voice features. These then correspond to physiological characteristics such as the rigidity and control of the vocal folds and vocal apparatus generally. When compared to say, free speech analysis, this technique offers increased consistency and comparability across patient demographics. Of the phonemes, the vowel /a/ (as in 'cat') is the most widely used. This is mainly due to the more distinct formants (harmonic frequencies) that compose this phoneme and that the produced sound is less affected by the particular physiology of the subject's vocal apparatus beyond the vocal folds as seen with phoneme types such as fricative and nasal. Distinct formants are preferable as a number of vocal characteristic calculations are based on the F_0, or fundamental frequency, of the phonation.

Submitted 14th July 2017. This research is part of a collaboration between University of Oxford and Hoffman-La Roche with funding provided by the EPSRC, MRC and Roche. The work was conducted as a rotation project in the Computer Intelligence in Biomedical Monitoring group of Oxford University's Institute of Biomedical Engineering, as part of the Doctoral Training Centre's Systems Approaches to Biomedical Sciences DPhil Program.

S. Bostock is with the University of Oxford, Doctoral Training Centre, Oxford, UK (email: stefan.bostock@spc.ox.ac.uk)

M. de Vos is with the University of Oxford, Institute of Biomedical Engineering, Oxford, UK

D. Wolf, F. Lipsemeier and M. Lindemann are with Hoffman-La Roche, pRED division, Basel, Switzerland.

Previous studies have shown this technique can be used to both accurately classify PD patients from healthy controls as well as UPDRS scoring within 7.5 points of clinical estimates [13]. However this was done in a controlled environment with negligible background noise and using the speech pathology specific Intel At-Home Testing Device (AHTD) high quality recording device. For the true benefits of such remote telemonitoring to be realized, the recordings must be made on commercially available devices in real-world environments. In this paper, we investigate the feasibility of this idealized scenario by analyzing 6655 voice recording from 56 people with Parkinson's (PWP) and 37 healthy controls, collected as part of a recent PD clinical trial. Additionally, we suggest a robust methodology for standardizing the feature analysis of phonations, partially based on investigating the effects of noise and analysis window parameters on the voice features.

II. Data

A. Overview

The data used in this research is in the form of audio recordings of PD patients and healthy controls (HC). The patients were part of a clinical trial of Prothena's PRX002 drug candidate; being developed as part of a partnership with Roche since December 2013 [14]. The healthy controls were recruited and tested by Roche independently of the trial for comparison to the patient samples.

B. The trial

The Roche-Prothena collaborative clinical trial identified as NCT02157714 was a Phase 1b study entitled '*A randomized, double-blind, placebo-controlled, multiple ascending dose study of PRX002 administered by intravenous infusion in patients with Parkinson's Disease*'. The purpose of this trial was to assess the safety, tolerability and pharmacokinetic profile of the drug across various dosing volumes, however UPDRS was also monitored [15].

The drug is a disease modifying humanized monoclonal antibody that targets α-synuclein, a key constituent of Lewy bodies, themselves a characteristic symptom of PD and specific forms of dementia [16]. It has been shown in animal models that α-synuclein specific antibodies help the clearance of the accumulated extracellular protein deposits and a reduction in neuropathology [17]. It is thus hoped that such a mechanism will translate into the clinic and as such is a prime candidate for PD immunotherapy.

The patients were grouped into placebo and five different dosing cohorts and each received a single intravenous infusion. UPDRS assessments were generally conducted on initial screening, dosing (day 0), day 8 and day 64.

C. Data Collection

Each patient was provided with a Samsung S6 smartphone. On these phones, an application was installed that enabled the patient to run a battery of tests away from the clinic. The test that this paper analyses involved the patient being asked to produce the sustained phonation, /a/, as the phone recorded for 30 seconds.

The patients were recruited on the U.S. eastern seaboard and recorded samples whenever and wherever they chose to, between January 2014 and August 2016. The healthy controls however, were recruited from Basel, Switzerland and tested between June and September 2016 with a single UPDRS assessment on commencement. Subjects recorded different numbers of samples depending of their compliance with the smartphone tests and the longevity of trial period, ranging from 1 to 229. Unfortunately, a technical error resulted in 12 of the patients and their voice samples being unable to be paired with the patient information.

The smartphones include standard commercial microphones with a sampling frequency of 44.1 kHz and a bit rate of 64 kbps.

D. Demographics

A breakdown of the demographics and how the samples are distributed across these is shown in Table I. It is noteworthy that for both HC and PD, males numbers outweigh female and that PD generally record more samples than HC.

TABLE I
SUBJECT DEMOGRAPHICS & DISTRIBUTION

Subject Group	Subject Gender	Subject Count	Sample Count	Initial Age		Initial UPDRS	
				Mean	Stdev	Mean	Stdev
HC	M	28	1187	57.25	7.61	3.61	2.78
	F	9	355	52.22	8.70	1.44	1.24
PD	M	36	4018	58.86	8.14	45.08	16.71
	F	8	824	52.50	8.47	37.75	19.00
	?	12	265	?	?	?	?

The UPDRS and age are taken as the values recorded on subject screening, as these change throughout the trial. Due to data collection error, twelve patients demographic and UPDRS data is unknown.

III. Methods

The overall aim of the work is to investigate whether feature analysis of the speech samples can be used in predicting subject classification (PD versus HC) or staging (UPDRS score). An overview of the investigations overall methodology is shown in Fig. 1 and a graphical explanation of some of the sample processing steps in Fig. 2.

A. *Sample Pre-Processing*

For this dataset, the first second of each sample needs to be ignored even if it includes a subject's phonation. This is because the phone vibrates on recording commencement to inform the subject to start. This translates to two loud 'beeps' as seen in Fig (2a).

B. *Longest Phonation Identification*

Regardless of the instructions that accompany the smartphone application task, many of the 30 second recordings include multiple utterances and as such a single phonation needs to be selected and its location in the recording identified. This is also required for a standardized phonation analysis workflow as future trials may have different instructions. Although unfounded, it was decided that the most robust option would be to select the longest identifiable phonation present in each recording. This is to ensure that the most complete and representative phonation is selected (Fig. 2d). A quick and effective method for this was adapted from a silence removal and speech segmentation implementation developed by Giannakopoulos in 2010 [18].

This algorithm calculates the signal energy and spectral centroid ('centre of gravity' of the signals frequency

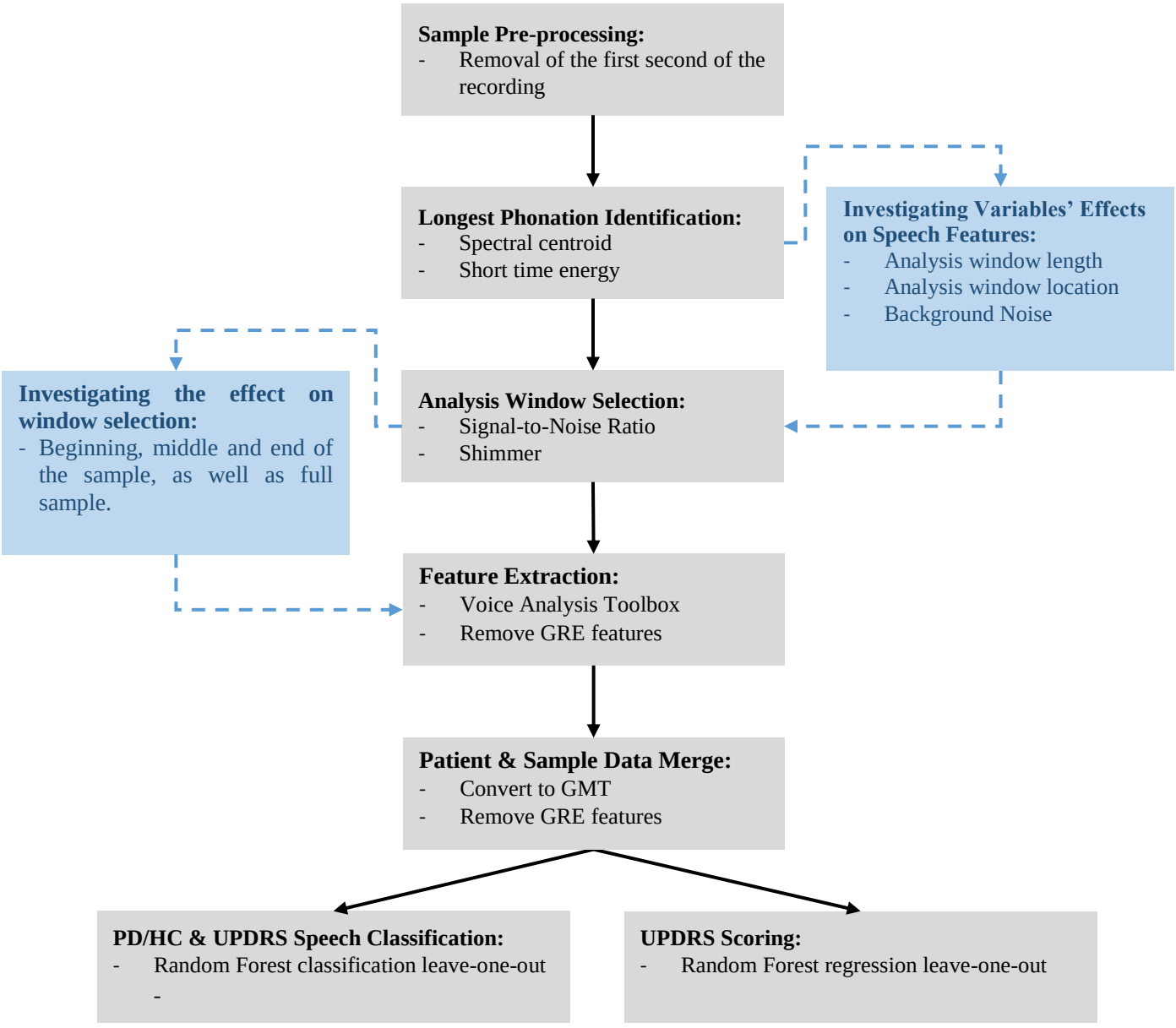

Figure 1: Overview of the investigative route and proposed workflow. Grey boxes are the suggested methodology for consistent voice analysis with key points on what contributes to that step. Blue boxes are the investigations that took place and informed the next steps as well as produced results, however they did not require any additional methodological changes to conduct.

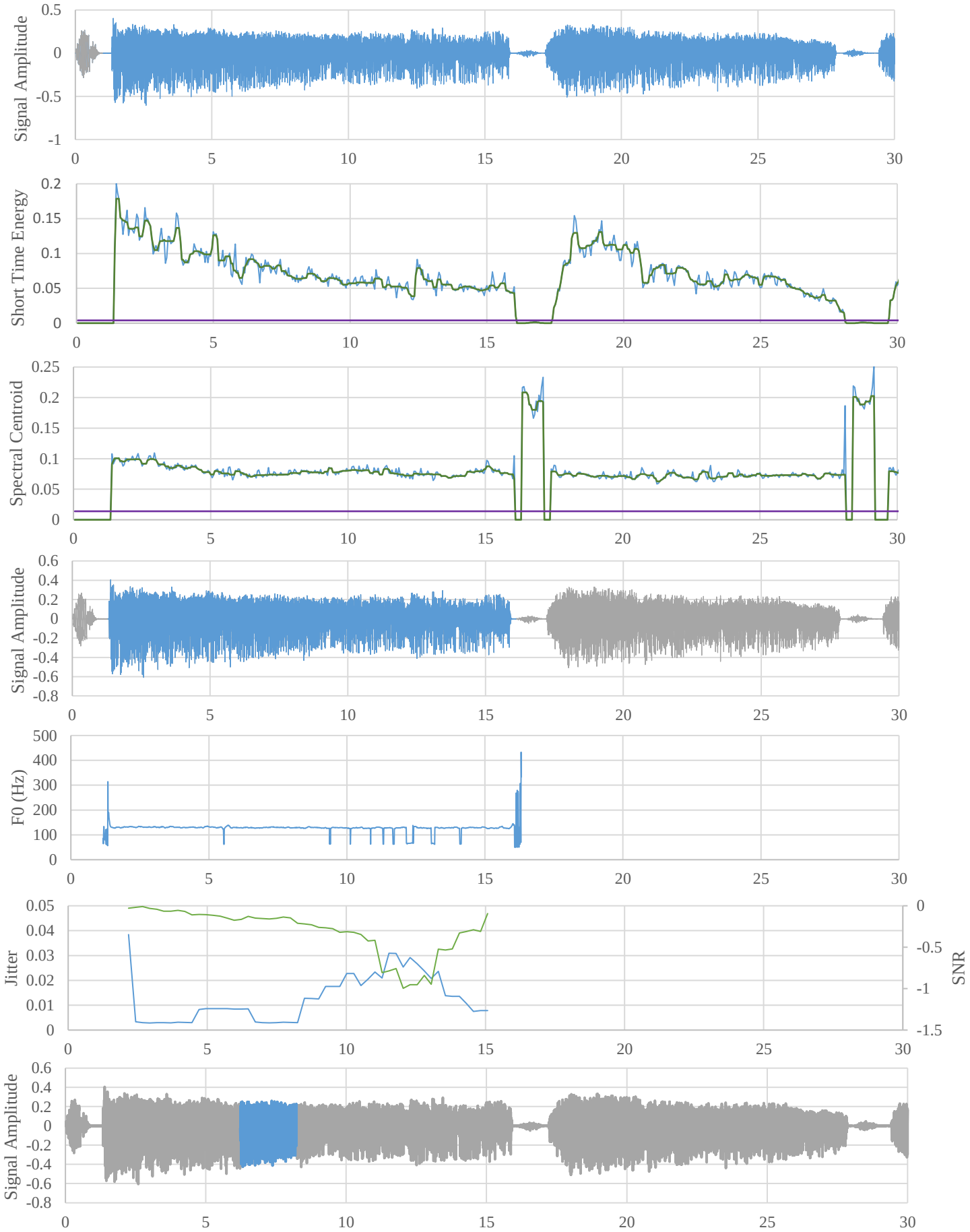
Signal Amplitude
Short Time Energy
Spectral Centroid
Signal Amplitude
F0 (Hz)
Jitter
SNR
Signal Amplitude

Figure 2. (previous page) Graphical overview of the newly proposed analysis window identification methodology. (a) The first second of the signal is removed from consideration to eliminate the two beeps (which can be seen here) which inform the user of the commencement of the recording. (b) The short time energy is calculated for 50ms non-overlapping frames and then smoothed and a threshold calculated (blue, green and purple respectively). (c) The spectral centroid is calculated for 50ms non-overlapping frames and then smoothed and a threshold calculated (same colouring). (d) The spectral centroid and short time energy signals are compared to their thresholds and for periods where both exceed their thresholds, a signal is identified. The longest continuous signal is considered to be the longest phonation and the rest of the recording removed from consideration. (e) The F_0 is calculated using the SWIPE algorithm. (f) Jitter (blue line) and the relative signal comparison (SNR, green line) for 2 second overlapping windows every 0.5 seconds. (f) The lowest five values of jitter are identified, and the one with the lowest respective SNR value is selected, the final 2 second window for analysis is that centring on this selection.

spectrum) for non-overlapping 50ms frames across the full recording (Fig. 2b and 2c). These features are good at isolating regions of speech when background noise is low compared to the signal noise and when background noise is of a spectral composition dissimilar to the phonation. Both of which are likely as the phone is held close to the subjects face and the likelihood of background noise similar to sustained phonation is negligible. These values are then smoothed.

When both of these features is over a threshold value -a function of the local maxima of the features- the signal is considered to be speech (Fig 2a, 2b and 2c).

The justification of the next step, analysis window selection (the region of the phonation used to calculate speed features), was informed by three side investigations the basic methodology of which is as follows. Firstly, the effects of the analysis window length was tested by calculating the features of a number of samples as the window length was extended from the first 0.5 seconds to 10 seconds in 0.5 second increments. Secondly, evidence suggests that the signal will change across the phonation, such as reduced volume or stability as the subject tires. This was investigated by sliding a fixed, 2 second, length window over a series of phonations. Finally, the effects of background noise on the extracted features was investigated by recording phonations with increased volume of background noise and inspecting feature variation. The background noises used were that of restaurants. This was chosen as it includes a number of common types of noise (as ascertained by manually listening to many of the samples) such as music, environmental noise, multiple and single people talking.

C. Analysis Window Selection

Using a small window rather than the whole phonation is suggested for several reasons: consistency, reproducibility, reduction in the chances of including background noise events, and to reduce processing times. The last point is crucial as the time taken to extract the features from the selected region of phonation is exponentially proportional to the analysis window length (see Fig. 3).

The newly proposed window selection method involves firstly evaluating the F_0 of the signal using the SWIPE algorithm [19]. From this a measure of its stability, jitter (Table II), can be calculated (Fig. 2e) whereby lower values indicate a smaller change in frequency over time. Both jitter and the F_0 are obtained using Mike Brookes' Matlab VOICEBOX toolbox [20]. Additionally, a signal-to-noise ratio (SNR) function is applied to each frame, comparing it to the first frame. This should provide an indication of the amount of additional frequency elements that are in that frame compared with the first. The idea being that if background noise or microphone static occurs in that frame a higher value will be obtained (Fig. 2f). We consider F_0 stability to be the most important element (partly as noise destabilizes it regardless as seen in the results) and as such select the frames with the five lowest jitter values and from these the single frame with the lowest SNR as our final selection.

In addition to this newly proposed window selection method, we also extracted the first, last and middle 2 seconds of the phonation, as well as it in its entirety, for comparison of its effectiveness. Beyond having been used in previous studies, each of these has justification. The beginning of the phonation is often the point of greatest volume, however stability has often yet to be established. The middle is likely the most stable part of the utterance, however consistency may be sacrificed as the amount of effort and volume throughout the phonation is probably not linear and different between subjects and even recordings. The end would potentially accentuate any characteristics of PD as the subject tires, however F_0 is often unstable at this point. Finally, the full sample would likely be more representative, however it has a much higher chance of containing background noise and also makes the problem almost computationally intractable (Fig. 3).

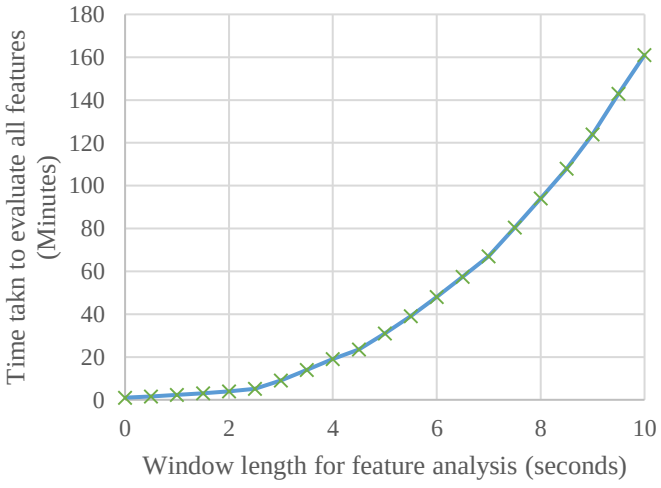

Figure 3. Time taken to perform all feature calculations by the length of the signal provided. The computer used to obtain these figures was running an Intel® Core™ i7-4770 CPU at 3.40GHz, with a RAM of 32Gb and a 64-bit Windows operating system.

D. Feature Extraction

With an analysis window identified, 339 features that quantify a range of speech characteristics are extracted using the Voice Analysis Matlab toolbox developed by Tsanas et al. [21], [22]. These predominantly focus on fundamental frequency and amplitude as well as the proportion of the signal that is attributed to noise and its consistency (Table II). In PD patients, decreased muscle control and vocal fold rigidity lead to variations in these values when compared to healthy subjects [21].

Due to time taken to calculate the entire feature set (Fig. 3), break down of how each feature group attributed to this time was conducted (Fig. 4). As EMD-ER contributes only 7 features, but the majority of the calculation time, it was excluded from consideration in the categorization and regression methods.

E. Patient & Sample Data Join

With each sample having 332 features evaluated, this data was joined with the patient data where available: UPDRS score, gender and age. For each patient there were up to 5 UPDRS scores taken over the length of the trial. As such those samples taken before the first or after the last assessment were given these values. Those between assessments where evaluated as if linear change had occurred

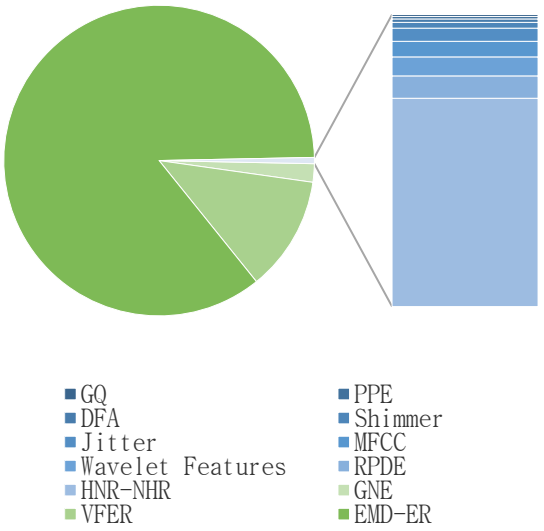

Figure 4. Relative time taken to extract each group of features. Determined using a clean sample from a healthy control, a 2 second analysis window and the newly proposed window selection methodology.

TABLE II
OVERVIEW OF SPEECH FEATURES

Feature Name	Count	Feature Number	Description
Jitter	22	1 to 22	Variation in fundamental frequency over time, clearly seen to be exacerbated in PD
Shimmer	22	23 to 44	Amplitude instability over time. PD exhibits increased instability.
Harmonics to Noise Ratio (**HNR-NHR**)	4	45 to 48	Ratio of how much of the signal can be attributed to harmonics and how much is turbulent airflow or noise
Glottal Quotient (**GQ**)	3	49 to 51	Quantification of the cycles in glottis opening and closing
Glottal-to-Noise Excitation (**GNE**)	6	52 to 57	Measure of noise in speech using nonlinear energy concepts
Vocal Fold Excitation Ratio (**VFER**)	7	58 to 64	Amount of noise in speech as determined using linear, nonlinear and entropic energy techniques
Empirical Mode Decomposition Excitation Ratio (**EMD-ER**)	7	65 to 71	Signal to noise ratio based on EMD-based energy
Mel Frequency Cepstral Coefficients (**MFCC**)	83	72 to 154	Metric of the short-term power spectrum of sound
Wavelet Features	182	155 to 336	Variations in F0
Pitch Period Entropy (**PPE**)	1	337	Instability in pitch
Detrended Fluctuation Analysis (**DFA**)	1	338	The stability and self-similarity of noise in speech
Recurrence Period Density Entropy (**RPDE**)	1	339	Measure of the consistency of vocal fold periodicity

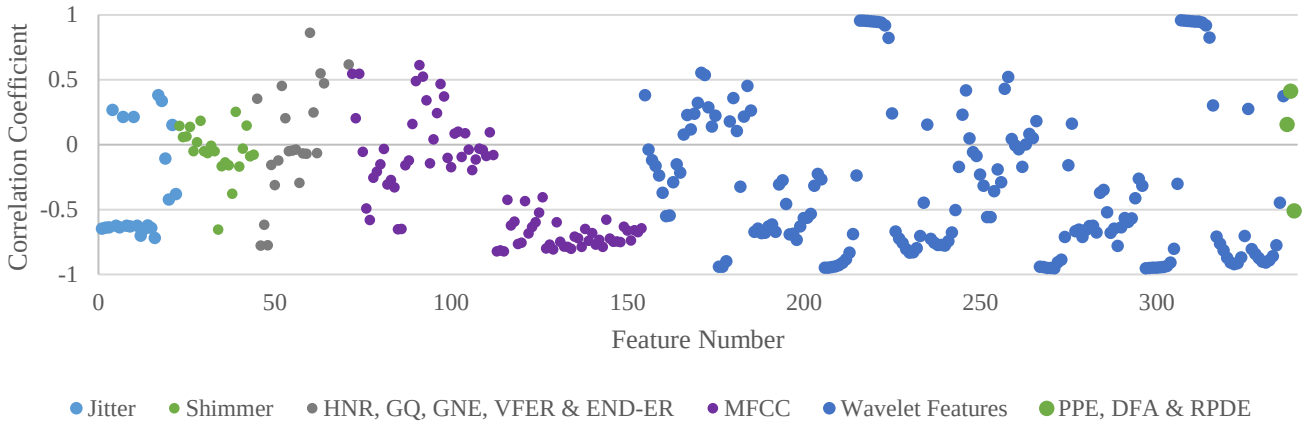

Figure 5. Correlation between analysis window length and 339 speech features. The Pearson correlation coefficients are coloured by the feature type. It is noteworthy that a number of patterns can be observed: many of the jitter features appear to be consistently correlated, as is MFCC, and the curves towards +1 and -1 in the wavelet features and VFER suggest a more complex correlation.

between the tests.

Each sample also had a date-timestamp. These were extracted and converted to GMT for consistency across the trial sites. The percentage of the way through the day was then calculated for each sample and added as an additional feature, as peoples voices have a propensity to vary according to circadian rhythms [23].

F. Classification and Regression

In order to assess the ability to determine people with PD from HC an unpublished random forest (RF) classification algorithm was used whereby each subject was tested against a model trained on all the other subjects. This replace-one strategy was considered the fairest method of testing. For UPDRS scoring a RF regression algorithm was used. Each training session was comprised of 1000 trees with a bag size of 12. RF were considered a suitable machine learning technique as Breiman demonstrates that they generally avoid overfitting which would have been of special concern with a data set of moderately few subjects and a range of sample numbers per subject [24].

IV. Results

A. Analysis Window Length

The linear correlation between analysis window length and each feature was assessed (Fig. 5). Although features may relate to analysis window length by a more complex relationship, due to the number of features, this was deemed an efficient overview. Indeed, there are some clear patterns in these values such as the consistent negative correlation in a number of the jitter as well as MFCC and VFER features. Additionally, the curves that can be observed towards the +1 and -1 values in the MFCC feature group suggest a more complex correlation. On closer inspection, the relationship is clear, an example of just three features is shown in Fig. 6,

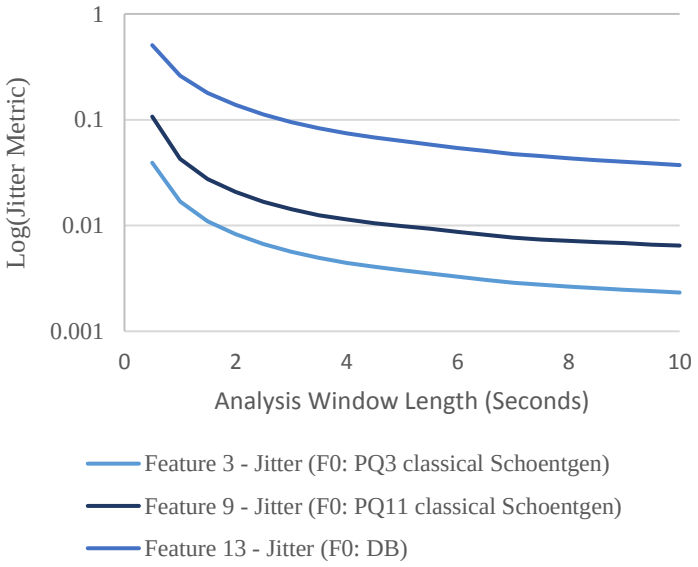

Figure 6. Variation of jitter features with analysis window length. Each of the aforementioned patterns (Fig. 5) was noticed to have a similar relationship to that graphed here as an example.

however all of these follow a similar pattern. Additionally, very few features have a value around 0. Combined, these observations suggest that analysis window length is intrinsically linked to most of the feature values, no doubt by the method by which they are calculated.

B. Analysis Window Location

Qualitative observation of the change in feature values across each sample confirms that much variation is seen throughout, an example of five features can be seen in Fig. 7. This is as expected, and is at least in part, some of the motivation behind the study and features in the first place. What is notable in Fig.7 is that the times at which variation occurs is not consistent, and therefore each feature must be picking up a different real-world variation. Therefore, what

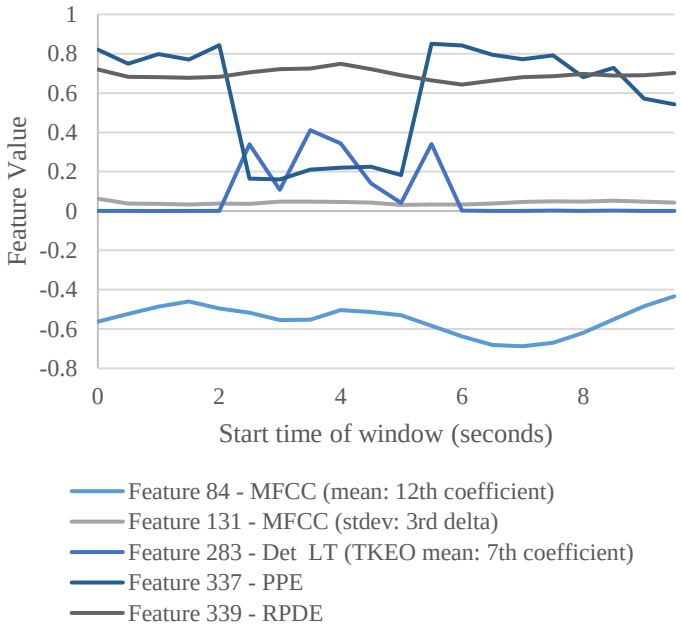

Figure 7. Variation of five features across 10 seconds of a phonation. The phonation used was from a single HC with negligible background noise.

is not clear is whether these variations are caused by changed in the speech, for example the subject running out of breath and straining the phonation, or external factors, such as change in the distance of the phone from the subjects mouth or background noise.

C. Background Noise

Similarly to analysis window length, the Pearson's correlation coefficient was calculated for each features response to increasing volume in background noise (Fig. 8). With the exception of the wavelet features, it appears that there is little pattern in the response to increased background noise. Interestingly however, jitter appears to have little linear correlation with noise. On closer inspection there is large variation in jitter, however it seems random. This is expected as the fundamental frequency of diverse background noise is highly variable. Shimmer seems to consistently increase with noise, as expected, as the volume of the background noise overwhelms the subject. It is worth noting however, that qualitatively, the volume of the background noise was very high. When recording, the top 70% of the recordings where at a background noise volume level substantially higher than both what would be a sensible expectation of the subjects recording environment and what was qualitatively observed on listening to many of the samples.

D. Classification and Regression

The overall accuracy (proportion of all samples correctly classified as PD or HC), subject accuracy (the proportion of subjects that have a majority correct sample classification), specificity, sensitivity, and counts for true positive, true negative, false positive and false negative were calculated for each of the sample window types (Table II). From this we can see that there is a slight advantage to using the entire phonation. However, considering the computation time comparison (~200 vs ~20 hours) it appears that the suggest method for feature extraction is a good approximation. It must be mentioned for transparency however that values in Tables III, IV and V and the hierarchy of accuracy in Table III does change a little depending on the parameters used and also due to the eponymous randomization process of in-bag feature selection involved in most implementations of random forests. Additionally, for this model only the 81 subjects with full patient information were used. When the full dataset was used classification accuracy dropped. This is

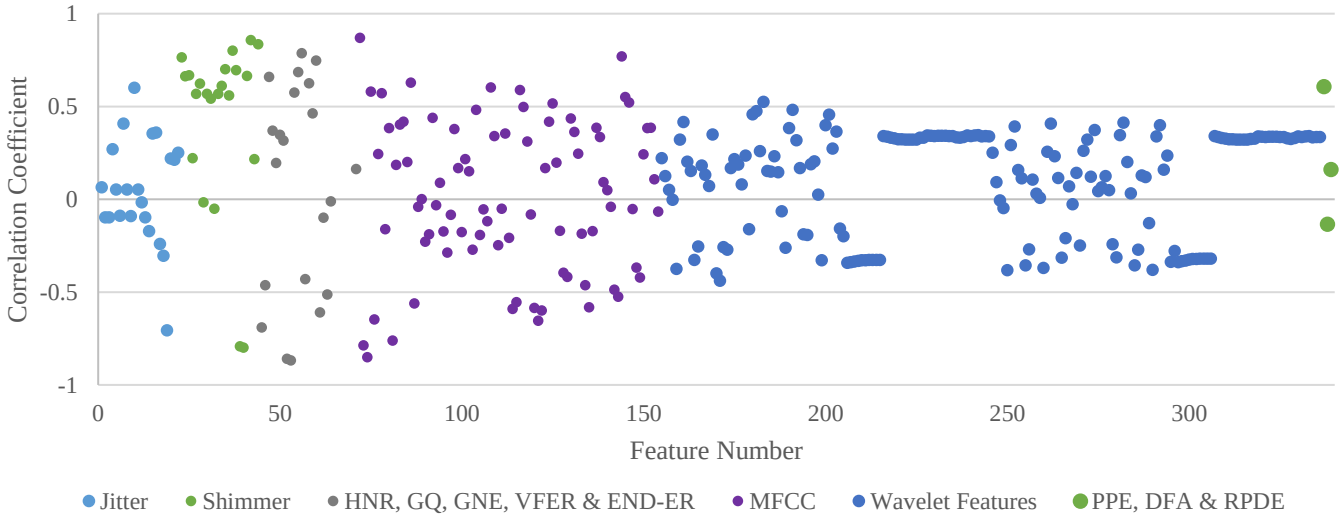

Figure 8. Correlation between background noise volume and 339 speech features. The Pearson correlation coefficients are coloured by the feature type. With the exception of the wavelet features and jitter there seems to be limited correlation, however, values are generally quite high.

TABLE III
RANDOM FOREST PD VS HC CATAGORISATION RESULTS FOR DIFFERENT ANALYSIS WINDOWS

Analysis Window Type	Sample Accuracy	Subject Accuracy	Subject Sensitivity	Subject Specificity	Subject True Positives	Subject False Positives	Subject True Negatives	Subject False Negatives
Beginning 2s	67.0 ± 3.5%	75.0%	95.5%	64.3%	21	1	27	15
Middle 2s	64.7 ± 3.9%	67.2%	85.7%	58.1%	18	3	25	18
End 2s	66.6 ± 3.8%	71.9%	84.6%	63.2%	22	4	24	14
Proposed 2s	68.6 ± 3.7%	78.1%	95.8%	67.5%	23	1	27	13
Full Sample	70.3 ± 3.7%	78.1%	92.3%	68.4%	24	2	26	12

TABLE IV
RANDOM FOREST PD VS HC CATAGORISATION RESULTS FOR DIFFERENT ANALYSIS WINDOWS

Subject Group	Sample Accuracy	Subject Accuracy	Subject Sensitivity	Subject Specificity	Subject True Positives	Subject False Positives	Subject True Negatives	Subject False Negatives
Male Only	70.3 ± 3.6%	78.1%	66.7%	92.9%	24	12	26	2
Female Only	58.8 ± 9.5%	64.7%	37.5%	88.9%	3	5	8	1
All Subjects	67.6 ± 3.4%	71.60%	54.5%	91.9%	24	20	34	3

TABLE V
RANDOM FOREST UPDRS SCORING ERROR

Subject Group	Mean Sample UPDRS Error	Mean Subject UPDRS Error
Male Only	12.34 ± 0.15	13.15 ± 1.77
Female Only	22.28 ± 0.30	21.48 ± 3.25
All Subjects	13.78 ± 0.15	14.38 ± 1.60

TABLE VI
FEATURE IMPORTANCE RANKING

Feature	Mean Relative Importance
Proportion of Day Passed	41.35 + 0.36
Age	36.24 + 0.47
MFCC (mean: 2nd coefficient)	34.61 + 0.31
Jitter (mean: F0 TKEO 5%)	33.74 + 0.23
Jitter (mean: F0 range 5 to 95%)	32.79 + 0.30
Entropy (det: log3 coefficient)	31.27 + 0.24
MFCC (mean: 0th coefficient)	29.38 + 0.26
Jitter (mean: F0 TKEO 25%)	24.01 + 0.22
Jitter (mean: F0 TKEO 95%)	22.13 + 0.22
MFCC (mean: 7th coefficient)	19.42 + 0.17

probably because the addition of basic subject information such as age, time of day and gender are vital characteristics in the model. This is heavily supported by the identification of these as the most important features in Table VI. We can also see how the different analysis windows compare across the subject group in Fig. 10. From this (as well as observations during optimization) we can ascertain that there are several subjects that are always almost 100% inaccurately classified regardless of technique and optimization. Furthermore, it appears that the full phonation and proposed analysis window outperform other windows primarily on those subjects that occupy the ambiguous middle ground (45-65%).

In Tables III, IV and V, ‘subject’ rather than ‘samples’ accuracy is focused on. This is considered a reasonable thing to do as any real world implementation of such a methodology to diagnose and monitor PD could rationally require multiple recordings over a period of time; as day-to-day and hour-to-hour UPDRS and speech varies.

Table IV informs us that the model is in fact heavily affected by the gender of the subject. The disparity between male and female classification accuracies is almost certainly due to the comparative dearth of female subjects in this study.

We can also see from Table III and IV that specificity is generally a lot higher than sensitivity. This is likely due to the fact that, from a disease severity perspective, there is a

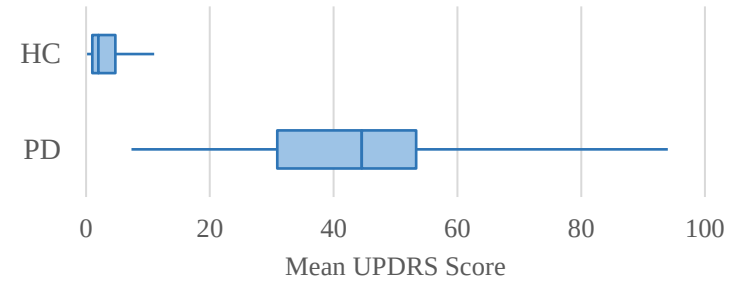

Figure 9. Distribution of subject UPDRS scores. In order to a provide a clearer overview than Table I

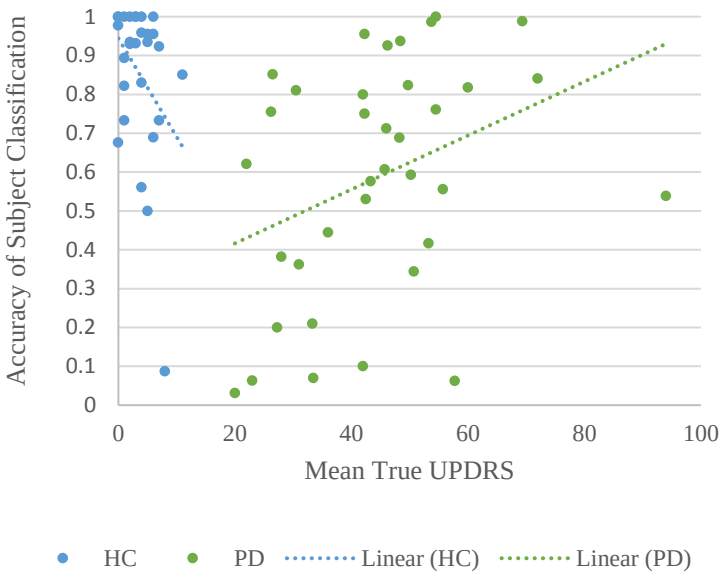

Figure 10. Distribution of subject mean rue UPDRS scores against their classification accuracies. The linear line of best fit is plotted for both HC and PD. It is clear from these that capacity to delineate the patient groups at the extremities is greater.

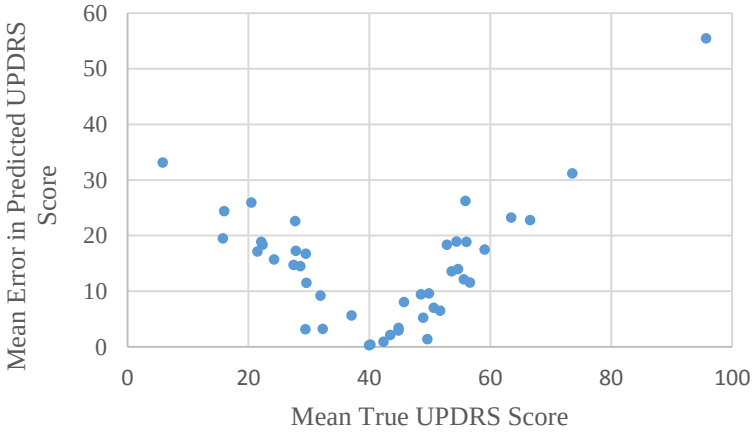

Figure 10. The RF Regression model error in predicting UPDRS score against the mean true score for each subject. Evidentially, the model is biased due to the disproportionate number of subjects with a score of 20 to 60.

much higher number of samples with a low UPDRS, whereas those subject with PD actually have a wide range of severities. As such it is easier for the model to identify what makes someone healthy. This is further suggested by Fig. 9 which gives a clearer overview of the distribution of UPDRS scores than Table I. Here it can be seen that some PD patients actually have lower UPDRS scores than HCs. This concept is heavily supported when one considers Fig. 10. Here we can clearly see greater capacity to identify PD at higher UPDRS scores, and increased ambiguity at the frontier between PD and HC.

Table VI shows the mean relative importance of the top ten most important features used for all sample classification. Most interesting is the fact that 'proportion of day passed' (essentially a percentage value of the time of day the recording was made) and the age of the subject were the most important. Age is somewhat unsurprising as the effect of age on speech is qualitatively observable in everyday life and has been quantified in academic research. To support the previous point on gender, age also affects men's and women's voice differently [25]–[28].

Despite the UPDRS RF regression model only taking into account PD patients, the capacity to predict UPDRS score was heavily insufficient for any useful applications such as patient monitoring (Table V). We believe that there are two key reasons for these results. Firstly due to time limitations the algorithm was barely optimized. However on inspection of individual sample results it appeared that there was a propensity for prediction to overestimate scores that should be lower than 40 and underestimate those over 40. This is confirmed by Fig. 11. The likelihood is that this is caused by the fact that the subjects tended to cluster around 40 and thus were not equally representative of the full range of potential UPDRS score. Fig. 9 shows that there were few subject over 55 and none over 94 when the maximum is 176.

V. Discussion

A. Analysis Window Length and Location

Our investigations dig deeper into feature variation due to window length and location than any other papers and show that it is indeed an important consideration. Therefore researchers should certainly note how they address these variables to ensure fair comparison between works.

With regard to window length the most important issue is that one is being consistent. It does however appear that many of the features tend towards stability the longer the window, however this must be balanced with computational tractability.

We have shown that analysis window location can have a substantial effect on classification results. In accordance with this we have suggested and demonstrated the efficacy of a more consistent methodology. This however should certainly be further investigated and improved upon as if researchers wish to avoid using the entire phonation, standardization of the analysis window selection method is crucial for the

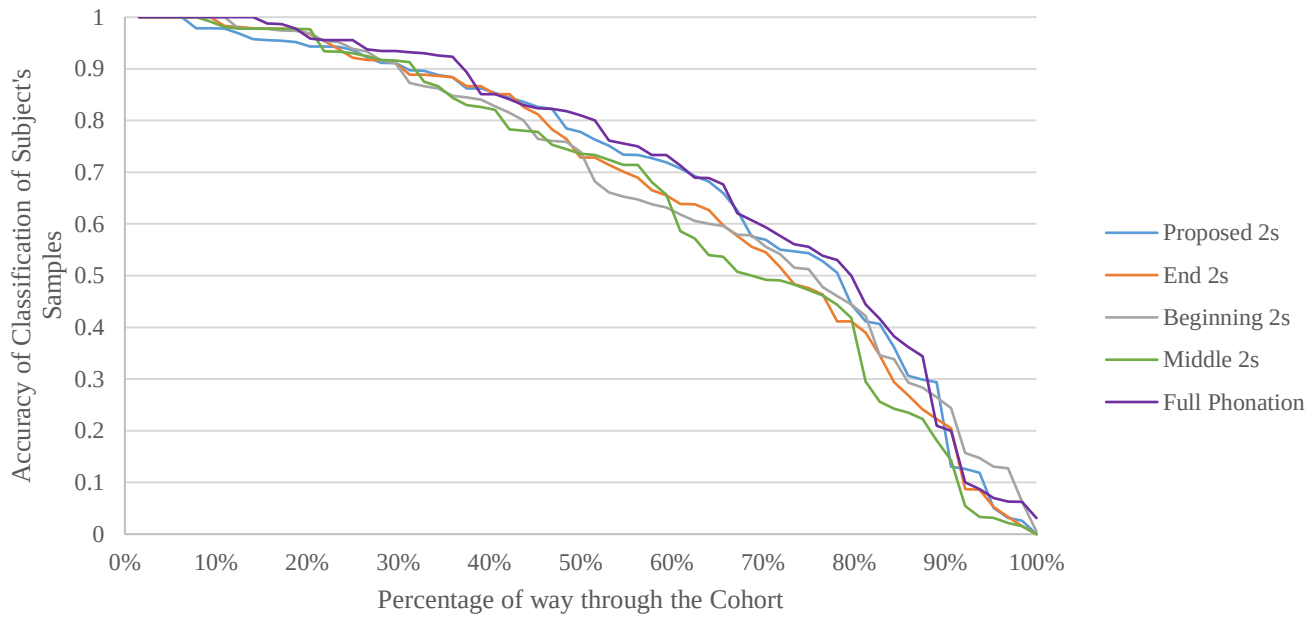

Fig. 12 Classification accuracy curves for each of the analysis window types. To achieve this, subjects were ranked and sorted by their classification accuracy and each subject evaluated as a proportion of the entire cohort.

credibility of the technology now that proof of concept has been achieved a number of times.

Using the full phonation does however seem to be a very effective method. This is most likely because it picks up the entire profile of the utterance, for example if a phonation is started at a high level of effort, it will account for this when it assess the sharper drop of in volume or stability towards the end.

B. Background Noise

We found that our results regarding the effect of background noise of speech features, especially wavelet features are similar to previous efforts [29]. There is clearly some effect of background noise on the feature values, therefore the most useful course of action is for the phone applications and clinicians to emphasise the importance of undertaking the recording in a quiet environment. Having said that, we believe that the overall effect on classification or UPDRS scoring is actually fairly minimal. From listening to the recordings, the vast majority of subjects did record in a quiet environment. Furthermore, it seems that for the effects of background noise to be substantial the volume has to be very high when compared to the phonation, which is unlikely, due to proximity of the phone to the subject's mouth. Additionally, if one uses the proposed analysis window selection methodology, noisy parts of the phonation are generally ignored due to the focus on low jitter values, or if the full phonation is used, then environmental sounds will likely only occur in a small proportion of the sample. It was noted however that F_0 was unstable when signal noise, such as microphone static was present. Therefore, further instruction on the subject not holding the phone too close to their face or speaking too loud would be recommended. The importance of this is noted by the fact that the previous studies have used data gathered using the state-of-the-art and bespoke Intel AHTD which includes a head mounted microphone maintained at around 5cm from the speakers mouth [13].

Our further testing on the importance of background noise would involve digitally adding increasing volumes (this should provide more reliable results) of different types of noise, as types such as musical, environmental, dialogue all have different effects [30]–[32].

C. PD vs HC Classification

One of the key findings is the effectiveness of adding both age and time of day as features in the model. We do not believe that this has been done before, and although we appreciate that it undermines the purity of only analyzing the signal, as well as raising issues with honest and accurate subject compliance outside of the clinic, we believe that the levity of the purpose will encourage users to enter such data correctly.

Overall, the results were comparable to previous studies using RF for classification [13], [33]. However, although very unlikely, we are unsure as to whether the drug used in the trial may direct effects on speech.

A major issue that we had was the disparity in accuracy of male and female cohorts. We believe that this is mainly due to the fact that there are far fewer women in the trial and their UPDRS is on average lower (Table I). It might be argued that a slight gender bias is preferable, however the gender prevalence of PD still a matter of debate. Regardless, future studies should include an even proportion of women, potentially by not using clinical trial subjects as there is a noted gender disparity in clinical trials [34], [35].

On a positive note, the subjects were well age matched within their own gender, although the men tended to be older

than the women, an issue that we have demonstrated as important.

Additionally, culture and language have a noted effect on voice characteristics and therefore there might be a geographic disparity contributing to the segregation of Swiss HCs and Americans with PD [36].

D. UPDRS Scoring

There was no discernible capacity to predict UPDRS accurately. This has been shown to be partly due to the relatively low and grouped UPDRS scores. However other studies have used subject groups with not dissimilar demographics, sample numbers and UPDRS scores [37], [13], [33]. Therefore, fundamentally, the model needs to be improved. We do hope however that the other findings may aid this.

E. Future Prospects

Further considerations such as finding a way for the subject to hit the same pitch, phoneme and volume is crucial. This would aid inter- and intra-subject correlation. For example, an interesting peculiarity of human behavior is that we tend to increase pitch with volume and how we do this depends on age and sex [38]. Additionally it was noted on listening to the samples that some subjects actually annunciated the phoneme /ʌ/ or /æ/ which have a fundamentally different F_0 and formant profile [39].

A greater cohort size and more longitudinal data would also enable the amelioration of subtle voice effects. For instance, voice features can fluctuate due to a number of variables outside the scope of this investigation; such as other pathologies, dialect, alcohol, smoking, environment, mood, facial expression and even personality or who the perceived listener is [40]–[48].

Overall we believe that this work adds to the corpora that suggests that speech analysis can be an effective, remote, non-invasive and inexpensive method for PD diagnosis and monitoring. This has been a strong step in showing that similar results can be obtained using smartphones.

The fact that the specificity is much higher than selectivity means that it would be a very suitable tool for clinical trial recruitment and may help to address the issue that around 15 to 30% of PD patients are undiagnosed [49], [50]; and of those that do, they get assessed by clinicians less than twice a year on average [51]. While also pushing down the vast costs of disease management ($316m in physician visits per annum in the U.S. alone - circa 2010 [52]). Perhaps, it may also enable the delineation of subtypes in what truly is a heterogeneous disease; with only 10% of cases attributed to genetic mutations and many with as broad a pathologies as oxidative stress, mitochondrial dysfunction and protein mishandling [53] .

VI. Conclusion

Overall, this research has confirmed that previous methods used to classify Parkinson's disease patients from healthy individuals using sustained phonation samples also works outside of a controlled environment, using smartphones to take the recording. Although classification accuracy was similar to those previously reported, there was a substantial reduction in the accuracy of UPDRS scoring using random forest regression models. However this is likely due to the unrepresentative distribution of UPDRS amongst the subject group as well as time limitations for project completion.

In addition to this affirmative confirmation we have investigated the dynamics of voice features in relation to: the length of the signal being analysed, which part of the signal is analysed and how much background noise is in the recording. From this we have ascertained that one must be careful when comparing papers on the subject, especially when such methodological characteristics are not noted. In addition we have proposed a moderately effective method for consistently extracting a useful segment of the phonation signal for analysis.

Finally, we must emphasize the importance of patient demographics, such as age, gender and UPDRS score distribution when constructing and testing such predictive models.

Acknowledgments

The work was support by Hoffman-La Roche with their provision of the samples from the clinical trial, and by their continued help and hosting at the pRED division in Basel.

References

[1] M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group.," *Neurology*, vol. 54, no. 11 Suppl 5, pp. S21–3, 2000.

[2] L. E. Clavería, J. Duarte, M. D. Sevillano, A. Pérez-Sempere, C. Cabezas, F. Rodríguez, and J. de Pedro-Cuesta, "Prevalence of Parkinson's disease in Cantalejo, Spain: a door-to-door survey.," *Mov. Disord.*, vol. 17, no. 2, pp. 242–9, Mar. 2002.

[3] Z.-X. Zhang, G. C. Roman, Z. Hong, C.-B. Wu, Q.-M. Qu, J.-B. Huang, B. Zhou, Z.-P. Geng, J.-X. Wu, H.-B. Wen, H. Zhao, and G. E. Zahner, "Parkinson's disease in China: prevalence in Beijing, Xian, and Shanghai," *Lancet*, vol. 365, no. 9459, pp. 595–597, Feb. 2005.

[4] B. S. Schoenberg, B. O. Osuntokun, A. O. Adeuja, O. Bademosi, V. Nottidge, D. W. Anderson, and A. F. Haerer, "Comparison of the prevalence of Parkinson's disease in black populations in the rural United States and in rural Nigeria: door-to-door community studies.," *Neurology*, vol. 38, no. 4, pp. 645–6, Apr. 1988.

[5] K. Wirdefeldt, H.-O. Adami, P. Cole, D. Trichopoulos, and J. Mandel, "Epidemiology and etiology of Parkinson's disease: a review of the evidence," *Eur. J. Epidemiol.*, vol. 26, no. S1, pp. 1–58, Jun. 2011.

[6] J. Fish, "Unified Parkinson's Disease Rating Scale," in *Encyclopedia of Clinical Neuropsychology*, New York, NY: Springer New York, 2011, pp. 2576–2577.

[7] D. A. Bennett, K. M. Shannon, L. A. Beckett, C. G. Goetz, and R. S. Wilson, "Metric properties of nurses' ratings of parkinsonian signs with a modified Unified Parkinson's Disease Rating Scale.," *Neurology*, vol. 49, no. 6, pp. 1580–7, Dec. 1997.

[8] R. Camicioli, S. J. Grossmann, P. S. Spencer, K. Hudnell, and W. K. Anger, "Discriminating mild parkinsonism: Methods for epidemiological research," *Mov. Disord.*, vol. 16, no. 1, pp. 33–40, Jan. 2001.

[9] M. Richards, K. Marder, L. Cote, and R. Mayeux, "Interrater reliability of the unified Parkinson's disease rating scale motor examination," *Mov. Disord.*, vol. 9, no. 1, pp. 89–91, Jan. 1994.

[10] A. Siderowf, M. McDermott, K. Kieburtz, K. Blindauer, S. Plumb, and I. Shoulson, "Test-Retest reliability of the Unified Parkinson's Disease Rating Scale in patients with early Parkinson's disease: Results from a multicenter clinical trial," *Mov. Disord.*, vol. 17, no. 4, pp. 758–763, Jul. 2002.

[11] L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey.," *Folia Phoniatr. Logop.*, vol. 46, no. 1, pp. 9–17, 1994.

[12] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients.," *J. Speech Hear. Disord.*, vol. 43, no. 1, pp. 47–57, Feb. 1978.

[13] A. Tsanas, M. A. Little, and P. E. Mcsharry, "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning," 2012.

[14] "United Therapeutics Announces Signing Of Agreement For New Remodulin Delivery System." [Online]. Available: http://ir.unither.com/releasedetail.cfm?ReleaseID=889184. [Accessed: 15-Feb-2015].

[15] D. B. Schenk, M. Koller, D. K. Ness, S. G. Griffith, M. Grundman, W. Zago, J. Soto, G. Atiee, S. Ostrowitzki, and G. G. Kinney, "First-in-human assessment of PRX002, an anti-α-synuclein monoclonal antibody, in healthy volunteers," *Mov. Disord.*, vol. 32, no. 2, pp. 211–218, Feb. 2017.

[16] M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, J. Q. Trojanowski, R. Jakes, and M. Goedert, "‖[alpha]‖-Synuclein in Lewy bodies," *Nature*, vol. 388, no. 6645, pp. 839–840, Aug. 1997.

[17] E.-J. Bae, H.-J. Lee, E. Rockenstein, D.-H. Ho, E.-B. Park, N.-Y. Yang, P. Desplats, E. Masliah, and S.-J. Lee, "Antibody-Aided Clearance of Extracellular -Synuclein Prevents Cell-to-Cell Aggregate Transmission," *J. Neurosci.*, vol. 32, no. 39, pp. 13454–13469, Sep. 2012.

[18] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals , implemented in Matlab," pp. 2–4, 2010.

[19] 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. 1638–1652, Sep. 2008.

[20] M. (Imperial C. L. Brookes, "Voicebox: Speech processing toolbox for matlab," 2006.

[21] A. Tsanas, M. A. Little, and P. E. Mcsharry, "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning," 2012.

[22] 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*, 2010.

[23] V. Jónsdottir, A. M. Laukkanen, and I. Siikki, "Changes in teachers' voice quality during a working day with and without electric sound amplification," *Folia Phoniatr. Logop.*, vol. 55, no. 5, pp. 267–280, 2003.

[24] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[25] L. A. Ramig and R. L. Ringel, "Effects of Physiological Aging on Selected Acoustic Characteristics of Voice," *J. Speech Lang. Hear. Res.*, vol. 26, no. 1, p. 22, Mar. 1983.

[26] S. N. Awan, "The aging female voice: Acoustic and respiratory data," *Clin. Linguist. Phon.*, vol. 20, no. 2–3, pp. 171–180, Jan. 2006.

[27] M. Brockmann, C. Storck, P. N. Carding, M. J. Drinnan, T. S., D. H., B. T., S. C., and H. K., "Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults," *J. Speech Lang. Hear. Res.*, vol. 51, no. 5, p. 1152, Oct. 2008.

[28] R. F. Orlikoff and J. C. Kahane, "Influence of mean sound pressure level on jitter and shimmer measures," *J. Voice*, vol. 5, no. 2, pp. 113–119, Jan. 1991.

[29] L. Bozhilova, M. De Vos, A. Tsanas, D. Wolf, F. Lipsmeier, and M. Lindemann, "Voice Impairment as a Biomarker for Parkinson's Disease," no. August, pp. 1–7, 2015.

[30] H.-M. Yang, Y.-J. Hsieh, and J.-L. Wu, "Speech Recognition Performance under Noisy Conditions of Children with Hearing Loss.," *Clin. Exp. Otorhinolaryngol.*, vol. 5 Suppl 1, no. Suppl 1, pp. S73–5, Apr. 2012.

[31] J. Meyer, L. Dentel, F. Meunier, F. Pellegrino, and N. Grimault, "Speech Recognition in Natural Background Noise," *PLoS One*, vol. 8, no. 11, p. e79279, Nov. 2013.

[32] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 851–854.

[33] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[34] V. H. Murthy, H. M. Krumholz, and C. P. Gross, "Participation in Cancer Clinical Trials," *JAMA*, vol. 291, no. 22, p. 2720, Jun. 2004.

[35] A. Heiat, C. P. Gross, H. M. Krumholz, G. PJ, F. L, R. N, and S. K, "Representation of the Elderly, Women, and Minorities in Heart Failure Clinical Trials," *Arch. Intern. Med.*, vol. 162, no. 15, pp. 1209–1216, Aug. 2002.

[36] R. van Bezooijen, "Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women," *Lang. Speech*, vol. 38, no. 3, pp. 253–265, Jul. 1995.

[37] 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*, 2010.

[38] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *http://dx.doi.org/10.1121/1.427150*, Aug. 1999.

[39] J. L. Flanagan, "A Difference Limen for Vowel Formant Frequency," *J. Acoust. Soc. Am.*, vol. 27, no. 3, pp. 613–617, May 1955.

[40] B. L. Brown, W. J. Strong, and A. C. Rencher, "Perceptions of personality from speech: effects of manipulations of acoustical parameters," *J. Acoust. Soc. Am.*, vol. 54, no. 1, pp. 29–35, Jul. 1973.

[41] C. A. Klofstad, R. C. Anderson, and S. Peters, "Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women," *Proc. R. Soc. London B Biol. Sci.*, vol. 279, no. 1738, 2012.

[42] C. Gobl, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1–2, pp. 189–212, Apr. 2003.

[43] H. R. Gilbert and G. G. Weismer, "The effects of smoking on the speaking fundamental frequency of adult women," *J. Psycholinguist. Res.*, vol. 3, no. 3, pp. 225–231, Jul. 1974.

[44] N. M. Docherty and N. M., "Cognitive Impairments and Disordered Speech in Schizophrenia: Thought Disorder, Disorganization, and Communication Failure Perspectives.," *J. Abnorm. Psychol.*, vol. 114, no. 2, pp. 269–278, 2005.

[45] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000.

[46] J. L. Cummings, A. Darkins, M. Mendez, M. A. Hill, and D. F. Benson, "Alzheimer's disease and Parkinson's disease: comparison of speech and language alterations.," *Neurology*, vol. 38, no. 5, pp. 680–4, May 1988.

[47] D. B. Pisoni and C. S. Martin, "Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analyses," *Alcohol. Clin. Exp. Res.*, vol. 13, no. 4, pp. 577–587, Aug. 1989.

[48] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Percept. Psychophys.*, vol. 27, no. 1, pp. 24–27, Jan. 1980.

[49] B. C. L. Lai, M. Schulzer, S. Marion, K. Teschke, and J. K. C. Tsui, "The prevalence of Parkinson's disease in British Columbia, Canada, estimated by using drug tracer methodology," *Parkinsonism Relat. Disord.*, vol. 9, no. 4, pp. 233–238, Mar. 2003.

[50] D. Twelves, K. S. M. Perkins, and C. Counsell, "Systematic review of incidence studies of Parkinson's disease," *Mov. Disord.*, vol. 18, no. 1, pp. 19–31, Jan. 2003.

[51] J. Lökk, "Lack of information and access to advanced treatment for Parkinson's disease patients.," *J. Multidiscip. Healthc.*, vol. 4, pp. 433–9, 2011.

[52] S. L. Kowal, T. M. Dall, R. Chakrabarti, M. V. Storm, and A. Jain, "The current and projected economic burden of Parkinson's disease in the United States," *Mov. Disord.*, vol. 28, no. 3, pp. 311–318, Mar. 2013.

[53] L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *Lancet Neurol.*, vol. 5, no. 6, pp. 525–535, Jun. 2006.