Predicting Severity of Parkinson’s Disease with Typing Behavior: A Machine Learning Approach

Tamar Schaap

Internal Supervisors:

Ben Harvey

Chris Janssen

External Supervisor:

Ricardo van Domburg

Universiteit Utrecht

Abstract

Parkinson’s disease (PD) is one of the most common neurological diseases in adults over the age of 65. Current monitoring of disease stage and progression consists of physician visits, which is inefficient and unreliable since symptom severity varies throughout the day. A more practical solution could include monitoring behavior throughout the day. Previous work has found that typing behavior is an accurate way to differentiate people with PD from those without PD. This finding leads to the idea that it may also be useful to use keyboard characteristics to detect PD severity. The current study examined whether this was possible. Additionally, it compared three different machine learning methods: logistic regression, *k*-nearest neighbors, and random forests. Finally, it examined how accuracy of the classification of PD severity differed when including increasing amounts of keystrokes (1000, 2000, and 5000 keystrokes). It was found that the random forests classifier could predict PD stage moderately well in the 5000-keystroke dataset. However, there was no further difference in model type or clear pattern to show that models become more accurate with increased amounts of keystrokes. This study is a first step in examining how computer behaviors might be able to be used for predicting and potentially monitoring PD patients’ disease stage.

Key words: Machine Learning, Parkinson’s Disease, Typing Behavior, Computer-Human Interaction

1. Introduction

Parkinson's Disease (PD) is a neurodegenerative disorder that affects roughly two percent of individuals globally over the age of 65, making it one of the most common neurodegenerative diseases in the elderly (Parkinson's Australia). The main course of the disease is associated with low levels of dopamine in the brain, resulting in loss of motor function; disease onset and progression results in symptoms such as tremors, rigidity, postural instability, memory disturbances, and executive dysfunction. Although there is no cure at this time, current treatment focuses on alleviating symptoms with pharmaceuticals that increase the amount of dopamine in the brain. In its early stages, PD may have a mild impact on function, where symptoms cause a modest impact on function (Goetz et al., 2008). Later, PD impact may be moderate; in this stage, symptoms are frequent and/or intense enough to impact motor function considerably, but not prevent it. Finally, when PD gets into severe stages, symptoms usually prevent motor function.

PD severity is often assessed with the use of the Unified Parkinson's Disease Rating Scale (UPDRS; Appendix A), which provides insight into motor evaluations of the patient and is scored by a clinician. This scale ranges from zero, which corresponds to a healthy state, to 176, which corresponds to severe affliction (Movement Disorder Society Task Force, 2003). The use of this evaluation is time-consuming and needs to be carried out by trained medical personnel. Monitoring and treatment costs can be incredibly expensive, with those who have moderate to severe cases costing roughly five times as much as those with mild cases (Bohingamu Mudiyanselage et al., 2017). Besides use of the UPDRS, other methods of monitoring motor fluctuations in patients include asking the patient to recall the number of hours of "ON" and "OFF" time. "ON" time refers to periods when medications are effective in relieving symptoms while "OFF" time refers to periods when symptoms are present. However, this method of monitoring patients is open to skewed perceptions of "ON" vs

"OFF" times since use of some antiparkinsonian medications (e.g., Levodopa) are shown to improve memory function during "ON" periods (Costa, Peppe, Brusa, & Caltagirone, 2008). This variation in memory function may distort a patient's recollection of these periods, implying that patients are potentially less able to give objective and realistic overviews during an appointment.

1.1. Problems with Current Monitoring of Disease Progression in PD

PD is one of the costliest neurological diseases, coming in at roughly $14.4 billion in the United States and €13.9 billion in Europe annually (Gustavsson et al., 2011; Olesen et al., 2012; Kowal et al., 2013), and is expected to increase in the proportion of affected individuals due to increasing life expectancies (Dorsey et al., 2007). Currently, a significant portion of annual costs for PD goes to unexpected hospitalizations and outpatient care (Andlin-Sobocki, Jonsson, Wittchen, & Olesen, 2005; von Campenhausen et al., 2011). Patients with PD schedule appointments with their neurologists on average every two to six months (von Campenhausen, 2011); however, there have been instances of 100% of patients going untreated for long periods of time in rural areas (Willis et al., 2011). Even in urban areas, patients can go untreated roughly 40 – 50% of the time due to inadequate access to specialized care (Willis et al., 2011). Currently, patients might just be asked to check in with their neurologists to have them examine how they are doing and if they are managing their symptoms as expected. These appointments, however, do not have a high level of efficacy in monitoring patients due to appointment times being too short to capture the variation in symptom presentation during the day (Patel et al., 2017).

Ineffective monitoring of symptom progression and presentation can be extremely detrimental for patients, as they depend upon treatment for their prolonged independence. For example, if a patient has lapsed into a state where they may need extra care in the home or a different medication plan since they have worsened symptoms, this may go unnoticed if they are having an "ON" moment. Early detection of disease progression and treatment adjustment could lead to early intervention and treatment, resulting in an increase in the amount of time patients live independently to an average of 15 years as compared to 8 years with poor or late treatment (Men, 2013). This calls attention to an unmet need for clinical specialists and patients with PD: an inexpensive method of remote monitoring. It is possible that with improved remote patient monitoring, appointment necessity will be reduced, therefore reducing frequency of scheduled visits or resulting in more e-consultations. Additionally, improved remote monitoring systems could help reach hard-to-reach patient populations such as in rural areas. These remote monitoring systems can be incorporated into future e-Health solutions, which would ultimately help in relieving many outpatient costs (e.g., early recognition of advancing disease stage/symptoms or reduction of the number of annual physician appointments).

Machine learning offers the possibility of doing exactly this using different kinds of data. The reason why machine learning might outperform other methods of monitoring include its ability to discover hidden relationships in data, its resistance to data errors, its ability to provide interactive feedback/monitoring summaries to patients and physicians, and its ability to scale to a very large number of patients (Brunato, Battiti, Pruitt, & Sartori, 2013). Below we will discuss a few existing studies using machine learning for remote monitoring and explain why the current study might be a particularly practical solution for remote monitoring.

1.2. Existing Remote Monitoring Possibilities

1.2.1. Voice Recognition

PD affects the speech of afflicted individuals since it involves complex motor coordination. For example, PD symptoms that are present in speech include reduced loudness, increased vocal tremors, and breathiness. Dysphonia (inability to produce normal vocal sounds) and dysarthria (difficulty in pronouncing words) both relate to PD and have been used in speech recordings to distinguish people with PD from healthy controls (Little et al., 2011). The fact that voice characteristics can be used to make this accurate classification led to the assumption that it could also be used to classify the severity of PD in individuals. However, in a study by Bayestehtashk and colleagues (2015), tasks eliciting speech for PD severity assessment resulted in only moderately accurate PD severity discrimination; therefore, although this may have practical application in combination with other measures in the future, it is currently not reliable as a sole measure for disease severity.

1.2.2. Wearable Sensors and smartphones

In the last decade, wearable sensors have emerged as tools to continuously obtain information from PD patients in day-to-day life. Wearable sensors can be worn for extended periods on parts of the body. These kinds of sensors include accelerometers, gyroscopes, and vibration sensors and can be installed in devices such as fall detectors and mobility monitors. These kinds of wearable sensors have a lot of potential for accurately monitoring patients since they can continuously monitor a patient's symptoms within their homes throughout the day; however, they come with a number of drawbacks. First, since PD is more common in the elderly, this is not a very practical method for people who are unfamiliar or uncomfortable with technology; this has been documented as a difficulty and reason for not wanting to use them by this population (Kubota, Chen, & Little, 2016; de Lima et al., 2016). Second, patients can feel anxious or uncomfortable wearing these devices due to the appearance or fit of bulky sensors (Cancela et al., 2014), and there is a need for the development of unobtrusive monitoring systems for inside and outside of the home (Espay et al., 2016). In addition to this, wearable sensors can confuse some tasks like mowing the lawn with symptoms such as tremors and does not accurately assess gait in patients who live in cluttered environments

(Kubota et al., 2016); they may also not be able to distinguish slowness of movement due to bradykinesia versus fatigue or other factors (Espay et al., 2016). Finally, these kinds of wearable sensors and accompanying software are often very expensive, costing over $3000 excluding software (Zhang et al., 2016). This can easily become a major hurdle to people, especially to those who’s insurance will not help cover costs.

Many of these issues can be solved with the use of quasi-wearables, which can be used to monitor patients without other wearable devices, such as smartphones. Smartphones are not uncomfortable or intrusive; furthermore, they are relatively inexpensive when compared to larger wearable sensors. Smartphones can be used to install monitoring programs such as HopkinsPD (Zhan et al., 2016) and Apple ResearchKit (Neto et al., 2016). These kinds of applications often require patients to complete multiple rounds of active tests each day to monitor symptoms. This is done by administering several short tests such as ones to measure gait, posture, voice, balance, dexterity, reaction time, tapping, and memory. Currently, however, these smart phone applications have only been used to detect medication responses in patients with PD. Additionally, there are still drawbacks to using smartphones for PD monitoring purposes. First, requiring multiple tests to be taken each day by people who may have memory disturbances is not practical or reliable and would likely result in inconsistent usage. Relatedly, there is a high dropout rate from quasi-wearable applications, resulting in inconsistent or rare usage (Ledger D, McCaffrey D, 2014), likely due to the lack of meaningful feedback given by these applications (Espay et al., 2016). Both of these facts indicate that the use of a smartphone application would not be ideal for monitoring programs. An additional concern surrounding these systems is that invasions of privacy often accompany quasi-wearable sensors. With GPS being monitored, patient’s physical locations and homes can be located; this requires secure anonymization precautions in order to protect the privacy of patients.

1.3. Leads for New Programs

Although existing approaches might not be practical to implement, recent studies revealed a promising behavior that might be able to be used to predict PD severity—typing. Using typing behavior is more practical and low-cost than other proposed solutions since it is a common behavior that people engage in; in older individuals (over the age of 50), it was found in a survey that most people use computers, especially for word processing and emailing (Goodman, Syme, & Eisma, 2003). Through the use of keyboard recordings, individuals would not have to remember to take tests multiple times during the day or learn how to use a new technology, as typing on a keyboard is already familiar to most people. Keyboard recording software would be able to record the typing characteristics, and machine learning could be used to determine the severity of PD based on these characteristics.

Typing behavior can be highly individual and can be used to identify users (Das, Mukhopadhyay & Bhattacharya, 2014; Bartmann, Bakdi, & Achatz, 2007; Giot & Rosenberger, 2012; Kang, Choi, Kim, Ma, & Lee, 2015; Teh, Teoh, Tee & Ong, 2011). There are multiple factors that are affected by PD that would, in turn, affect typing behavior. For example, rest tremor is a highly recognized and common symptom of PD (Jankovic, 2008). Bradykinesia, or a slowness of voluntary movement, is another common symptom of PD that could likely affect typing behavior. Additionally, there is the unilaterality that is associated with PD, especially at its earlier stages (Jankovic, 2008). Finally, the rigidity that is seen in PD is common in the wrists (Jankovic, 2008), and could affect typing behavior. These symptoms could all be manifested in more or less of the following typing characteristics (depending on severity): reaction speed, hold times, latency times, asymmetry of responses between left and right hands, hesitations and pauses, and overall variability of movement. Recently, Adams (2017) used these features in a machine learning approach to distinguish individuals with early-stage PD to those who were neurotypical; in this study, he achieved 100% accuracy when using individuals with at least 2000 keystrokes.

Using these same features, it can be reasoned that this approach could also provide some insight into the severity stage of PD. For example, the unilaterality would likely decrease as PD progresses (Hoehn & Yahr, 1967), and the other symptoms (e.g., bradykinesia) could likely get worse, affecting aspects of typing behavior more. The current study will examine if characteristics of typing behavior can be used to predict the severity of PD in individuals.

Although Adams (2017) included all individuals who had typed more than 2000 keystrokes, it would be useful to know how many keystrokes can provide adequate accuracy into the severity of an individual's PD. By understanding how many keystrokes are necessary to reveal information about PD stage, we can create a threshold of keystrokes necessary for accurate machine learning models. Therefore, the current study will also be comparing the accuracies given for machine learning in datasets comprised of differing keystroke counts. This is important to investigate if this is to be used for diagnostic purposes, since it would be vital to know how many keystrokes are needed to give an accurate prediction of PD stage. Therefore, we will compare a dataset comprised of 1000 keystrokes, 2000 keystrokes, and 5000 keystrokes to determine how many keystrokes may be necessary to accurately predict PD stage.

Finally, the current study will investigate which kind of machine learning algorithm fits best with this kind of data. We will compare three different commonly used algorithms also used in the Adams (2017) study since all models used in his study resulted in accurate classification: multinomial ordinal logistic regression (LR), *k*-nearest neighbors (kNN), and random forests (RF). Since these are common algorithms, they would be easy to implement into any monitoring systems without deep understanding and heavy customization of models.

1.4. Research Questions

The three exploratory research questions posed in this study are the following: 1) can typing behavior predict PD severity, 2) which machine learning method best predicts severity, and 3) does accuracy of severity classification significantly increase with more keystrokes?

2. Materials and Methods

2.1. Participants

The study by Adams (2017) was approved by the Human Research Ethics Committee (in agreement with National Health and Medical Research Council guidelines; Australian Research Council) at Charles Sturt University (protocol number H17013). Participants were 50 to 80 years of age with a median birth year of 1948 and a median diagnosis year for those with PD of 2014. Participants were recruited with the help of Google Ads and the Michael J Fox Foundation "Trial Finder" facility. Participants could visit the research website to apply to partake in the research project, and inclusion criteria for participants were that they had to own and regularly use (at least a few minutes each day) a Windows computer and be at least 50 years old. Exclusion criteria were if they were diagnosed with any other neurological disorders.

2.2. Methods

Keyboard recordings were collected between July 2016 and March 2017 after participants downloaded and installed Tappy, a custom keylogging application, on their home computers. Participants were simply asked to continue using the computer as they normally would. Tappy ran in the background so that participants did not notice it running, and it recorded timestamps (with a timing accuracy within 3 milliseconds) when letter keys were pressed with specifications about whether the key was on the right (R) or left (L) side of the keyboard or if the space bar (S) was pressed. Participants were asked to ensure that they were the only user of the Window's account on which Tappy was installed. PD information was collected from participants by means of a questionnaire, which was distributed after the installation of Tappy.

Individual keystroke timings were saved and added to a CSV file on the participant's computer, which was automatically uploaded to the lab's database server once a day. In order to protect participants' privacy their personal information (e.g., names) was replaced by a randomized 10-digit code. In order to ensure what information was sent to the database server, participants were able to access their own data at any time.

2.3. Preparation for Data Analysis

2.3.1. Classification methods and metrics

We compared three different supervised machine learning algorithms (LR, kNN, and RF) which were also used in Adams's (2017) study in order to determine which algorithm works best to classify PD stages. All models were designed to maximize Cohen's Kappa, which is a performance metric that compensates for random hits and is particularly good for multi-class problems and imbalanced datasets (Ben-David, 2008a; Ben David, 2008b). The different machine learning algorithms are briefly described below:

> LR: LR is a supervised method that can be used for multi-class classification. We used ordinal LR with LASSO penalty in order to further reduce the effect of multicollinearity since LR is particularly sensitive to correlated features (Hosmer & Lemeshow, 1989; Ryan, 1997). Ordinal LR models can be used when the response variable belongs to one of multiple ordered categories.

kNN: This algorithm classifies unlabeled data based on their similarities with examples in the training set. The only adjustable parameter in this form of classification is *k*, the number of nearest neighbors to include in the estimate of class membership. It finds the *k* closest features in the training set and assigns to the class that appears most frequently within the *k*-subset. By varying *k*, the model can be made more or less flexible.

RF: RF is a classifier that consists of a combination of tree classifiers where each tree depends on the values generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector (Breiman, 1999). All trees have the same distributions (Breinman, 2001). RF models are relatively robust to outliers and noise in the data; therefore, they tend to do well in high-dimensional data (Breinman, 2001).

There are many different kinds of metrics that can be used to evaluate the performance of a model. The current study uses 2 different families of metrics—threshold and rank. Threshold metrics are used to minimize the number of errors a model makes, making them popular in different applications of classifiers. Additionally, we make use of rank metrics, which are important for cases in which good class separation is crucial. Several studies have shown that performance metrics often do not agree on a best model (Huang Lu, & Ling, 2003; Szöllősi et al., 2012; Duan et al., 2014). This is especially the case when working with imbalanced datasets and multiclass problems (Ferri Hernández-Orallo, & Modroiu., 2009); in these cases, more than balanced and binary cases, the different performance metrics tell us different things. In this study, we utilized three performance metrics in order to simplify the comparison/generalization to other studies; the results of the classifications were evaluated on the criteria of Accuracy, Cohen's Kappa, and Area Under the Receiver Operating Characteristic Curve (AUC). Each one is briefly described below.

Accuracy is a widely used threshold-based metric to determine class discrimination ability of classifiers. It is calculated from the confusion matrix by comparing the correctly classified cases to all cases. The biggest advantage is its simplicity; however, Accuracy as a metric can be misleading in imbalanced cases. For example, a classification method could have classified all cases into the majority class, in these cases, we know that the algorithm did not learn as we expected, but the Accuracy alone would not be able to provide insight into its poor performance. Additionally, Accuracy does not take the impact of chance into account (Powers, 2012). For these reasons, Cohen’s Kappa and AUC might provide more insight. Although Accuracy may not be the best metric for the current study, it is used in order to generalize better to other studies since Accuracy is the most widely-used performance metric for machine learning (Szöllősi et al., 2012).

Cohen’s Kappa is another threshold-based metric that takes into account the random correct classification and gives a “chance corrected coefficient of agreement” (Reed, 2000). Essentially, Kappa gives an approximation of how much better our classifier is performing over the performance of a classifier that guesses at random according to the frequency of each class. It is calculated as the difference between correct observations and expected outcomes, divided by the complement of expected outcomes. Cohen’s Kappa ranges from -1 to 1. According to Landis and Koch (1977), less than 0 indicates no alignment, 0 – 0.20 is slight alignment, 0.21 – 0.40 is fair alignment, 0.41 – 0.60 is moderate alignment, 0.61 – 0.80 is substantial alignment, and 0.81 – 1 represents almost perfect alignment. Cohen’s Kappa is commonly used as a measure of classifiers in medicine and statistics (Kaymak, Ben-David, & Potharst, 2012). Therefore, it is useful in this study since it is likely to be used in other related studies, which in turn makes it simple to compare to other studies. This, in combination with its preference over Accuracy in cases with imbalanced data makes it a good option as a performance metric.

AUC is a rank-based metric that gives an overall evaluation about classification abilities of the models. It is independent of prevalence and is considered highly effective for scoring the performance of models with ordinal data (Allouche, Tsoar, & Kadmon, 2006). However, it can give misleading information if the ROC curves are crossing (Hand, 2009). It also uses different misclassification cost distributions for different classifiers (Hand, 2009). To calculate the AUC, the true positive rate (sensitivity) is plotted against the false positive rate (1.0 – specificity) as the threshold varies from 0 to 1, where 1 relates to no error, and .5 relates to random models. A good model will achieve a high true positive rate while the false positive rate is still relatively small. Since the current study deals with multiclass classification, AUC in this study will be calculated as outlined by Hand and Till (2001) when there are more than two classes included for classification; this computes the average of multiple ROC curves. There has been some controversy about the use of AUC in multiclass situations since this method of calculating AUC assesses the average ability of separating any pair of classes. Although a high AUC calculated this way means that a classifier is good at separating most pairs, it is still possible that some classes are harder to distinguish. However, it is used in the current study since it measures accuracy in a genuinely different way than any other performance metrics (Ferri, Hernández-Orallo., & Modroiu, 2009). Additionally, it is highly recommended in situations with imbalanced data (Ertekin, Huang, Bottou, & Giles, 2007). Despite these benefits of using AUC, it has been known to mask poor performance as well (Jeni, Cohn, & De La Torre, 2013). Therefore, we wanted to use this metric in combination with other commonly used metrics.

2.3.2. Features

Participants’ typing information was used to create aggregated features for hold time, latency time, and flight time to be used in the models. Hold time is defined as the time between the press and release of a key. Latency time is the time between a key press of one key to the key press of the next key. Finally, flight time refers to the time between releasing a key and pressing the next key. Several features were derived from each of these keystroke characteristics using the side of the keyboard as well as the direction from which the previous keystroke came. This meant that a R, L, S, R-R, R-L, R-S, L-R, L-L, L-S, S-S, S-R, and S-L version of each characteristic was created (e.g., L_Hold_Time, and S-R_Flight_Time). For each of these, a mean, skew, kurtosis, and standard deviation was created for each participant. For hold times, an additional difference score was created between R and L means for a measure of asymmetry. Similarly, for latency time and flight time, difference scores were created between R-R and L-L means, and R-L and L-R means to create measures of asymmetry.

Additionally, although all combinations of keystroke tuples were originally included, it was found that there were very few cases of participants using the space bar twice in a row; therefore, all features using S-S were dropped, leaving 138 features. However, this still meant that there were many features in comparison to the number of participants in each dataset (Table 1). Additionally, when examining a correlation matrix of these features, the data presented that many were highly correlated with each other. Since many machine learning algorithms are sensitive to high-dimensional data, redundant features, or multicollinearity (Kotsiantis, Zaharakis, & Pintelas, 2007), it was decided that feature selection would have to be used in order to reduce the amount of features.

Table 1

Descriptive Statistics of Participants

Characteristic		With Levodopa			Without Levodopa		
	Number of Keystrokes	1000	2000	5000	1000	2000	5000
Gender	Male	68	54	38	42	32	24
	Female	76	68	51	34	31	25
Severity	Control	35	29	20	35	29	20
	Mild	48	41	33	26	22	20
	Moderate	48	42	27	13	10	7
	Severe	13	10	9	2	2	2
Tremors	Yes	69	59	44	24	19	16
	No	75	63	45	52	44	33
Sidedness	Left	33	32	25	13	13	12
	Right	40	34	28	18	15	13
	None	71	56	36	45	35	24

In order to address this, prior to training models, a Genetic Algorithm (GA) was used to reduce the number of features to only salient features. The GA was developed based on evolutionary principles of natural selection (Goldberg, 1988), and is often used for feature reduction (Siedlecki & Sklansky, 1993; Roth & Levine, 1994; Yang & Honavar, 1998; Sun, Bebis, Yuan, & Louis, 2002; Zamalloa et al., 2008). In a comparison between Correlation, Recursive Feature Elimination, and GA as feature selection techniques on three separate high-dimensional datasets, GA was the most effective feature selection technique (Glander, 2017). Additionally, in a study by Zamalloa and colleagues (2008), GA performed better than Principal Component Analysis (PCA) on untransformed data; for these aforementioned reasons, the current study utilized a GA in place of PCA, which was the preferred feature selection method in Adams's (2017) study.

2.3.3. Pre-Processing

Data frames were made consisting of participants and their aggregated keystroke information. Some machine learning methods such as kNN cannot perform with missing values (Kotsiantis, Zaharakis, & Pintelas, 2007); in order to resolve conflicts resulting from missing values, all missing values were fit with mean imputation. Furthermore, all variables were scaled and centered to ensure that no features outweighed others solely due to differences in scaling and unit differences between features. Three datasets were created by using only the first 1000 keystrokes of every participant in one dataset, 2000 keystrokes in a second dataset, and 5000 keystrokes in a third dataset. Any hold times, latency times, or flight times that lasted more than 3500 milliseconds were considered intentional and subsequently removed.

Adams (2017) removed participants who used Levodopa from the analysis presumably due to its use greatly affecting the ability to detect differences in keystroke information between PD controls and mild PD. In order to determine whether participants using Levodopa should be excluded from the analysis, we examined whether the influence of Levodopa was indeed as strong as Adams predicted. In order to do this, we used the 5000-keystroke dataset. We then used all three machine learning methods that would be used in the study to determine whether it was possible to predict Levodopa-use based on keystroke information.

Each model was run 25 times and collected Accuracy, Cohen's Kappa, and AUC as performance metrics. Additionally, we ran a non-learning Null model 25 times that always predicted the majority class. A GA was used to select which features should be included in the models. Then we compared the three performance metrics using three separate Analyses of

Variance (ANOVA's) to examine whether the models could predict Levodopa-use better than the Null model (Table 2). Since all three ANOVA's showed that there was a significant main effect for model type, we used Tukey's post hoc tests to examine how the models differed from each other. Tukey's post hoc comparisons confirmed that all models performed significantly better than the Null model for all performance metrics (Table 3), meaning that Levodopa-use indeed greatly influenced keystroke information. For this reason, participants taking Levodopa were removed from the main analysis. This exclusion left very few people (*N* = 2; Table 1) in the severe condition; therefore, we only used participants from the control, mild, and moderate condition in the analyses. Due to the focus of the study being in predicting disease stage, we are mostly interested in accurately predicting the mild and moderate groups. With this in mind, it was necessary to determine whether disease stage could be predicted when all three classes were included in the data.

Table 2

ANOVA's Comparing Performance Metrics of Models to Examine whether Levodopa-use could be Predicted

Effect	*F* statistic	df	Error df	η^2	*p*-value
Accuracy					
Model	.486	3	96	.350	< .001*
Kappa					
Model	.969	3	96	.345	<.001*
AUC					
Model	.278	3	96	.344	< .001*

Note: * < .0001

Table 3

Means, Standard Deviations, and p-values for Tukey’s Pairwise Comparisons of Performance Metrics for Models Predicting Levodopa-use

	Descriptives		Tukey’s Post Hoc Comparisons		
Model	*M*	*SD*	kNN	RF	Null
Accuracy					
LR	.583	.115	.4311	.918	<.001*
kNN	.623	.089		.818	<.001*
RF	.600	.065			<.001*
Null	.448	.104			
Cohen’s Kappa					
LR	.207	.162	.589	.941	<.001*
kNN	.254	.174		.264	<.001*
RF	.185	.130			<.001*
Null	0	0			
AUC					
LR	.608	.0837	.444	.954	<.001*
kNN	.638	.098		.190	<.001*
RF	.597	.069			<.001*
Null	.500	0			

Note * < .0001

Initially, in order to determine whether all three classes (control, mild, and moderate) could be included in the models for accurate performance, all models were trained using stratified 5-fold cross-validation and run once. The performance metrics and confusion matrices of these models were inspected in order to determine whether the moderate class (the underrepresented class) could be correctly classified. If it were shown that the use of all three classes resulted in poor classification of the moderate class, then the included classes would be restricted to include only mild and moderate cases. For the purposes of this study, these are the most important classes to discriminate between since they are the classes with some degree of PD severity (while the control group was never diagnosed with PD). Final performance metrics for these models can be seen in Table 4 below. Although the performance metrics indicate that the use of all three classes might result in relatively accurate predictions (since chance level for Cohen’s Kappa is 0 and .5 for AUC), it should be noted that due to an underrepresentation of patients with moderate impact in all three datasets, none of the classifiers were good at classifying this level; this was determined by examining the confusion matrices, which showed that the moderate class was rarely predicted correctly, and it was never predicted correctly in the 5000-keystroke dataset (Appendix B, Appendix C, Appendix D).

Table 4

Results of Classifier Accuracy in All Datasets Using Control, Mild, and Moderate Classes

Classifier	Accuracy	Cohen’s Kappa	AUC
1000 keystrokes			
LR	.346	.066	.566
kNN	.500	.228	.550
RF	.539	.232	.630
2000 keystrokes			
LR	.318	.044	.559
kNN	.500	.237	.628
RF	.546	.209	.712
5000 keystrokes			
LR	.235	-.270	.661
kNN	.647	.400	.691
RF	.706	.500	.738

Due to the poor classification of the minority class when all classes were included, this step was repeated using only the mild and moderate classes in order to examine whether this would result in better classification of the moderate class. Since this class was still underrepresented in comparison to the mild class, the models were run using oversampling. The results for these models can be seen in Table 5. As can be seen from the performance metrics and confusion matrices (Appendix E, Appendix F, Appendix G), this resulted in improved classification of the moderate class. Therefore, only the mild and moderate classes were used to create the models for statistical comparisons that determined whether models performed above chance and which models were significantly better performers than others.

Table 5

Results of Classifier Accuracy in All Datasets Between Mild and Moderate Classes Using Oversampling

Classifier	Accuracy	Cohen's Kappa	AUC
1000 keystrokes			
LR	.5	-.0426	.4792
kNN	.5714	.0455	.5208
RF	.5714	.0455	.5208
2000 keystrokes			
LR	.5833	-.1538	.4375
kNN	.6667	.25	.625
RF	.75	.4	.6875
5000 keystrokes			
LR	.5	-.3158	.3571
kNN	.7	.2105	.5952
RF	.9	.7368	.8333

2.4. Data Analysis

Following the lead of Adams (2017), all three datasets were split into training and test sets with a 0.65:0.35 ratio. Data organization was completed using the Python programming language (Python Software Foundation) while model calculations were carried out in R project (R Core Team) with models being created using the *train* function in R’s caret package (Kuhn, 2017). All models were specified to maximize Cohen’s Kappa, which has been recommended for imbalanced datasets (Ben-David, 2008a; Ben-David, 2008b). Each model’s performance was evaluated using the validation set. All three aforementioned performance metrics (Accuracy, Cohen’s Kappa, and AUC) were used to evaluate a model’s ability to discriminate between stages of PD.

In order to investigate which model performed as the best classifier and whether classification improved significantly with increased amounts of keystrokes, models were trained using the training data as a single fold and then tested on the validation set. An additional non-learning Null model was created that always classified all observations into the majority class. This Null model helps determine whether models are predicting disease stage better than chance. Models were run 25 times for each model and keystroke set, resulting in a total of 300 observations. The three performance metrics for each model were saved into a separate dataset, which was used to run three separate ANOVA’s, one for each performance metric. Where the ANOVA’s were significant, Tukey’s post hoc tests were used to identify significant differences between models.

3. Results

In order to address our research questions, we ran three separate ANOVA’s in order to test for the effect of conditions on Accuracy, Cohen’s Kappa, and AUC. The predictors included in each analysis were Model Type and Number of Keystrokes; however, we also included the interaction term between these two predictors in case there was a more complex pattern in the data. The results of these three ANOVA's can be seen in Table 6 and means and standard deviations for all models can be seen in Table 7.

Table 6

Results of ANOVA's for Accuracy, Cohen's Kappa, and AUC

Effect	*F* statistic	df	Error df	η^2	*p*-value
Accuracy					
Model	11.782	3	288	.109	< .001***
Keystrokes	32.068	2	288	.182	< .001***
Model*Keystrokes	4.825	6	288	.091	< .001**
Kappa					
Model	6.956	3	288	.068	<.001**
Keystrokes	4.622	2	288	.031	.011*
Model*Keystrokes	4.492	6	288	.086	<.001**
AUC					
Model	15.146	3	288	.136	< .001***
Keystrokes	10.163	2	288	.066	< .001***
Model*Keystrokes	10.542	6	288	.180	< .001***

Note: * < .05, ** < .001, *** < .0001

Table 7

Means and Standard Deviations of Models' Performance Metrics

	Accuracy		Cohen's Kappa		AUC	
Models	*M*	*SD*	*M*	*SD*	*M*	*SD*
1000 Keystrokes						
LR	583	.139	.069	.241	.558	.118
kNN	.591	.100	.073	.198	.558	.093
RF	.586	.115	-.007	.232	.511	.123
Null	.664	.108	0	0	.500	0
2000 Keystrokes						
LR	.573	.132	-.027	.241	.486	.123
kNN	.610	.124	.134	.225	.589	.114
RF	.649	.105	.146	.203	.574	.091
Null	.677	.073	0	0	.500	0
5000 Keystrokes						
LR	.632	.135	.037	.262	.553	.153
kNN	.712	.127	.112	.312	.555	.152
RF	.864	.178	.510	.869	.800	.242
Null	.732	.099	0	0	.500	0

When using Accuracy as the performance metric, there is a significant main effect for Model Type, *F*(3, 288) = 11.782, *p* < .001, and Number of Keystrokes, *F*(2, 288) = 32.068, *p* < .001. Additionally, there is a significant interaction between Model Type and Number of Keystrokes, *F*(6, 288) = 4.825, *p* < .001.Due to the significant interaction, we used the Tukey post hoc test to compare all models and keystrokes to each other. It was found that the kNN classifier at 5000 keystrokes (*M* = .712, *SD* = .127) and the Null model at 5000 keystrokes (*M* = .732, *SD* = .099) had significantly higher Accuracy values (Figure 1) than the LR model at 1000 keystrokes (*M* = .583, *SD* = .139) and 2000 keystrokes (*M* = .573, *SD* = .132). They also performed better than the kNN at 1000 keystrokes (*M* = .591, *SD* = .100) and the RF at 1000 keystrokes (*M* = .586, *SD* = .115). Additionally, it was found that the Null model at 5000 keystrokes significantly outperformed the kNN model at 2000 keystrokes (*M* = .610, *SD* = .124). Finally, it was found that the RF classifier using 5000 keystrokes (*M* = .864, *SD* = .178) had significantly higher Accuracy values than all other models. All *p*-values for model comparisons for Accuracy can be seen in Table 8.

Table 8

Results of Tukey's Post Hoc Tests for Accuracy

	Tukey's HSD Comparisons											
Models	1000 Keystrokes				2000 Keystrokes				5000 Keystrokes			
	LR	kNN	RF	Null	LR	kNN	RF	Null	LR	kNN	RF	Null
1000 Keystrokes												
LR		.999	.999	.452.	.999	.999	.749	.225.	.959	.012*	<.001**	.001*
kNN			.999	.630	.999	.999	.882	.366	.991	.028*	<.001**	.004*
RF				.511	.999	.999	.799	.268	.973	.016*	<.001**	.002*
Null					.277	.919	.999	.999	.999	.963	<.001**	.710
2000 Keystrokes												
LR						.996	.558	.117	.868	.004*	<.001**	<.001*
kNN							.992	.731	.999	.125	<.001**	.023*
RF								.999	.999	.805	<.001**	.410
Null									.980	.997	<.001**	.908
5000 Keystrokes												
LR										.469	<.001**	.149
kNN											<.001*	.999
RF												.009*
Null												

Note * < .05, ** < .0001

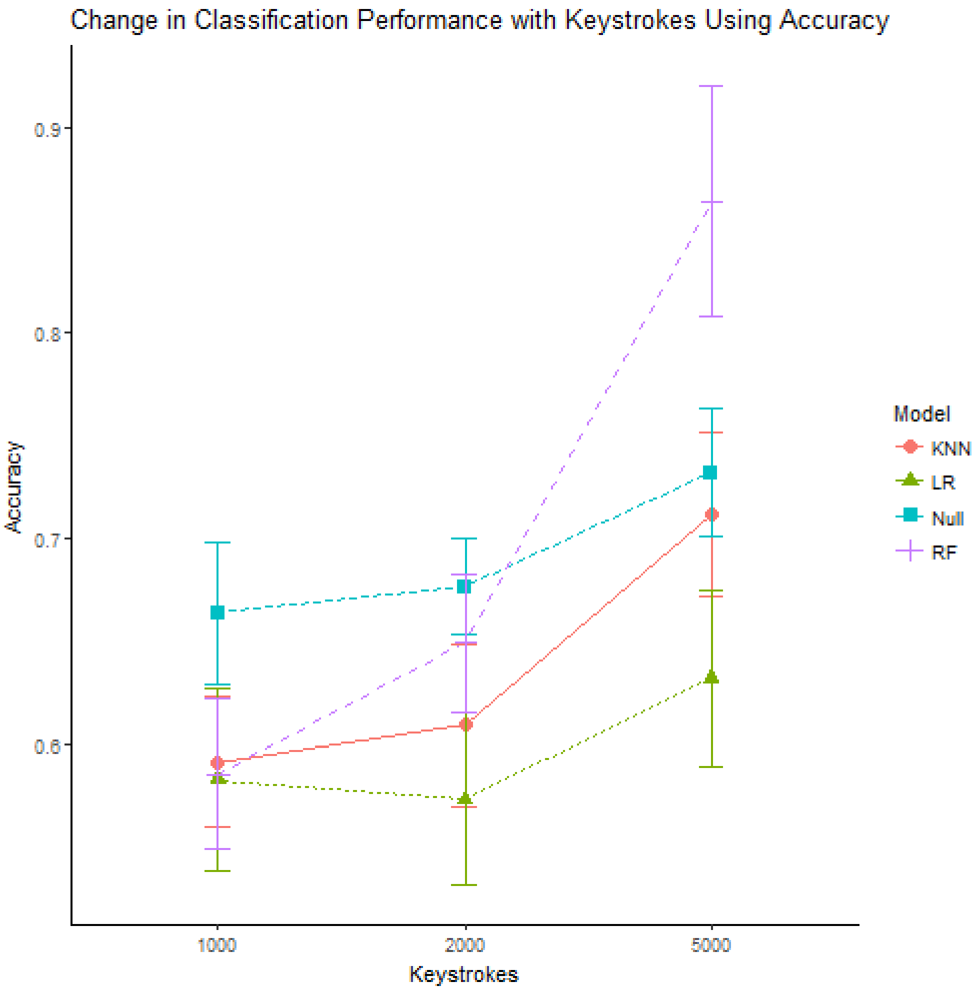


Figure 1.Interaction plot showing the means and standard errors of Accuracy for each model type at different amounts of included keystrokes. Here the kNN and Null model at 500 keystrokes outperformed the LR models at 1000 and 2000 keystrokes. They also outperformed the kNN and RF models at 1000 keystrokes. Additionally, the Null model at 5000 keystrokes outperformed the kNN model at 2000 keystrokes. Finally, the RF classifier at 5000 keystrokes significantly outperformed all other models.

When using Cohen's Kappa as the performance metric, there is a main effect for Model Type, *F*(3, 288) = 2.126, *p* < .001, and Number of Keystrokes, *F*(2, 288) = 4.622, *p* = .011. Additionally, there is a significant interaction between Model Type and Number of Keystrokes, *F*(6, 288) = 4.492, *p* < .001. When examining the interaction in closer detail with the use of Tukey's post hoc comparisons, it can seen that while the RF classifier (*M* = .510, *SD* = .869) using 5000 keystrokes was significantly higher (Figure 2) than all other models, other models did not differ significantly from each other or outperform the Null model. Since this was the only model significantly outperforming any other models, it was determined that it was driving the main effects found for individual predictors. All *p*-values for model comparisons for Cohen’s Kappa can be seen in Table 9.

Table 9

Results of Tukey’s Post Hoc Tests for Cohen’s Kappa

	Tukey’s HSD Comparisons											
Models	1000 Keystrokes				2000 Keystrokes				5000 Keystrokes			
	LR	kNN	RF	Null	LR	kNN	RF	Null	LR	kNN	RF	Null
1000 Keystrokes												
LR		.999	.999	.999.	.996	.999	.999	.999	.999	.999	<.001**	.999
kNN			.999	.999	.994	.999	.999	.999	.999	.999	<.001**	.999
RF				.999	.999	.922	.871	.999	.999	.977	<.001**	.999
Null					.999	.944.	.901	.999	.999	.985	<.001**	.999
2000 Keystrokes												
LR						.826	.749	.999	.999	.928	<.001**	.999
kNN							.999	.944	.996	.999	.002*	.944
RF								.901	.988	.999	.004*	.902
Null									.999	.985	<.001**	.999
5000 Keystrokes												
LR										.832	<.001**	.999
kNN											<.001*	.915
RF												.<.001**
Null												

Note * < .05, ** <0001

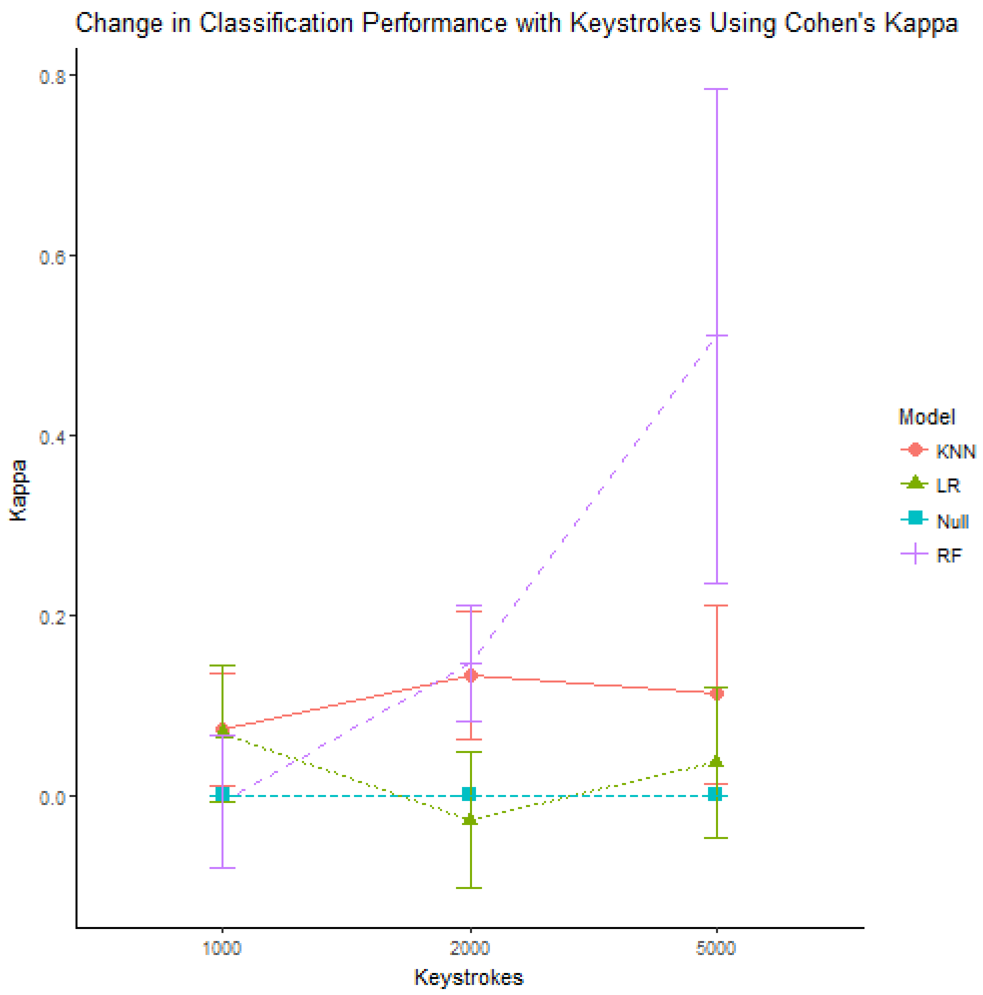


Figure 2. Interaction plot showing the means and standard errors of Cohen's Kappa for each model type at different amounts of keystrokes. Cohen's Kappa for the RF classifier at 5000 keystrokes was significantly higher than other models. A Cohen's Kappa value of 0 represents chance.

Finally, when using AUC as the performance metric, there is a main effect for Model Type, *F*(3, 288) = 15.146, *p* < .001, and Number of Keystrokes, *F*(2, 288) = 10.163, *p* < .001, Additionally, there is a significant interaction between Model Type and Number of Keystrokes, *F*(6, 288) = 10.542, *p* < .001. When using Tukey's post hoc comparisons to examine the interaction further, results showed that while the RF classifier (*M* = .800, *SD* = .242) using 5000 keystrokes was significantly higher (Figure 3) than all other models, all other models did not differ significantly from each other. In fact, no other models outperformed the Null model. Since the RF classifier at 5000 keystrokes was the only model significantly outperforming any other models, it was determined that it was driving the main effects found for individual predictors All *p*-values for model comparisons of AUC values can be seen in Table 10.

Table 10

Results of Tukey’s Post Hoc Tests for AUC

	Tukey’s HSD Comparisons											
Models	1000 Keystrokes				2000 Keystrokes				5000 Keystrokes			
	LR	kNN	RF	Null	LR	kNN	RF	Null	LR	kNN	RF	Null
1000 Keystrokes												
LR		.999	.971	.879.	.642	.999	.999	.879	.999	.999	<.001**	.879
kNN			.970	.875	.636	.999	.999	.875	.999	.999	<.001**	.875
RF				.999	.999	.513	.801	.999	.987	.983	<.001**	.999
Null					.999	.300	.588	.999	.928	.915	<.001**	.999
2000 Keystrokes												
LR						.122	.314	.999	.733	.707	<.001**	.999
kNN							.999	.300	.997	.998	<.001**	.300
RF								.588	.999	.999	<.001**	.588
Null									.928	.915	<.001**	.999
5000 Keystrokes												
LR										.999	<.001**	.928
kNN											<.001*	.999
RF												.009*
Null												

Note * < 001, ** < .0001

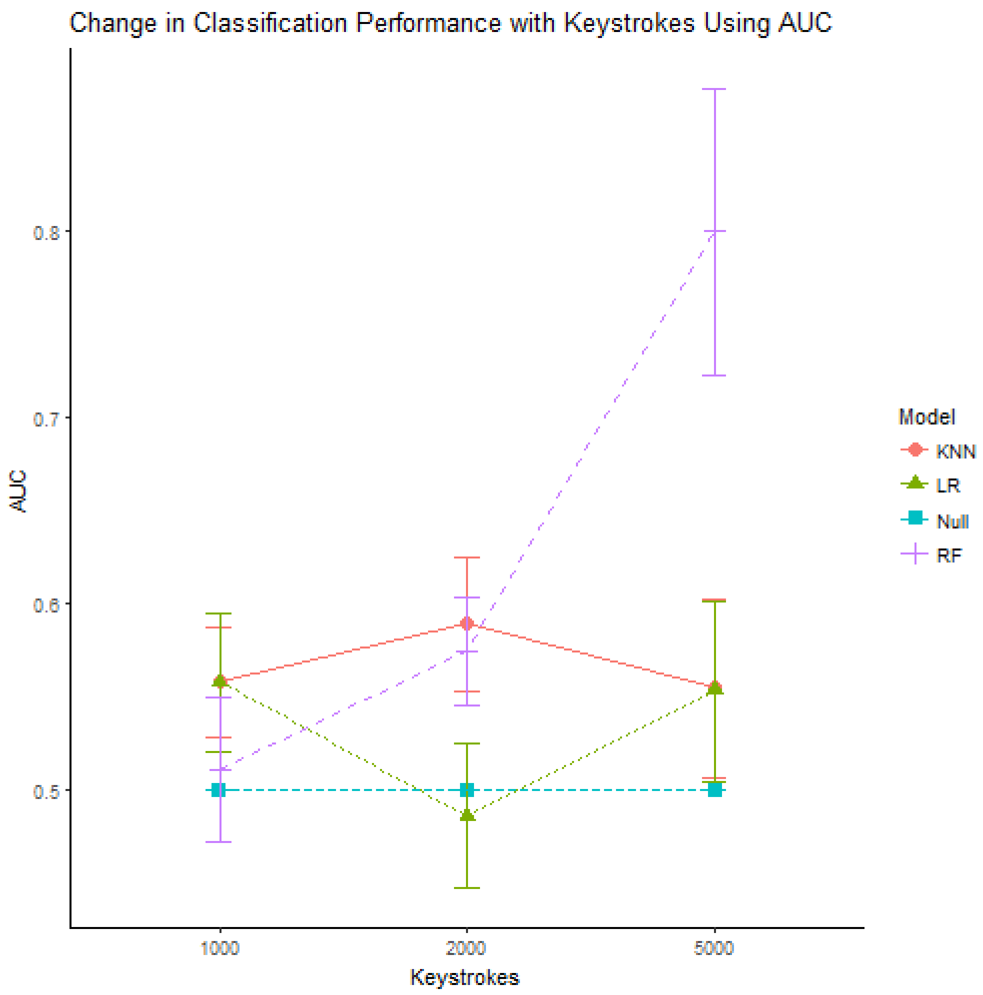


Figure 3. Interaction plot showing the means and standard errors of AUC for each model type at different amounts of keystrokes. AUC for the RF classifier at 5000 keystrokes was significantly higher than other models. An AUC value of .5 represents chance

In summary, RF appears to be the best performing classifier while using 5000 keystrokes; this holds for all three performance metrics. In addition, the kNN classifier at 5000 keystrokes does fairly well in comparison to most other models when using Accuracy as a performance metric. However, the Null model at 5000 keystrokes still outperforms it. Since the RF classifier at 5000 keystrokes is the only model that significantly outperforms the Null model using every performance metric, it is the only model that conclusively performs better than chance. Furthermore, there is no evidence that one model type outperforms other models consistently when controlling for number of keystrokes. Additionally, there is no evidence that models become significantly better performers as more keystrokes are used.

4. Discussion

The current study was conducted in order to answer three research questions: 1) can typing behavior predict PD severity, 2) which machine learning method best predicts severity, and 3) does accuracy of severity classification significantly increase with more keystrokes? In order to answer these questions, we used different machine learning methods on datasets comprised of differing numbers of keystrokes with the goal of determining whether keystrokes could be used to predict the stage of someone’s PD. Additionally, we compared LR, kNN, and RF as machine learning classifiers in order to determine which machine learning algorithm works best to correctly classify PD stages. Finally, we compared datasets composed of 1000 keystrokes, 2000 keystrokes, and 5000 keystrokes in order to determine whether increasing the number of keystrokes increased the classification performance.

In this study, it was hypothesized that keystroke characteristics could be used to determine PD severity. The current study has helped shed some light on this; it was found that typing behavior could moderately predict PD stage when using the RF classifier at 5000 keystrokes. According to Landis and Koch’s (1977) explanations of Cohen’s Kappa values, the RF classifier for the 5000-keystroke dataset performs moderately. Similarly, if using AUC as a metric, the RF classifier for the 5000-keystroke dataset performs moderately according to previous studies (Obuchowski, Lieber, & Wians, 2004; Mechref, Novotny, Kyselova, & Kang, 2008; Del Hoyo, Isabel, & Vega, 2011). With the use of both metrics, as well as the classifier being the only model to consistently outperform the Null models, it can be determined that the RF classifier at 5000 keystrokes performs moderately well and is the best-performing classifier. This is an important finding that has potential implications for the ways in which PD patients are monitored. Although performance metrics might have to be higher for direct implementation into monitoring programs, the finding that typing behavior carries information about disease stage in PD is promising for extended research into how closely inspecting computer interaction can reveal much about patient health.

The analyses were originally meant to include the data from the control group; however, machine learning with more than two classes in imbalanced datasets resulted in poor classification of the minority class. When examining the confusion matrices (Appendix B, Appendix C, Appendix D), it can be seen that while the control and mild classes were predicted with relative success, the moderate class was almost never predicted correctly. This finding is supported by previous studies, which have found that classification results from multimajority cases, in which there is more than one majority class (e.g., control and mild), are worse than other imbalanced cases due to the imbalance rate being severe. In these cases, oversampling also cannot be of use since it would cause overfitting to the minority class (Wang & Yao, 2012). Undersampling would also not work to help this since 1) the sample size is too small to consider dropping valuable learning opportunities in the training sets and 2) it would suffer in predicting majority classes (Drummond & Holte, 2003; Wang & Yao, 2012).

Although the RF classifier using 5000 keystrokes performed moderately well, RF classifiers at lesser keystrokes do not seem to perform significantly above chance. Additionally, the rest of the tested models also do not predict PD stage significantly better than chance. Therefore, it cannot be concluded that one type of classifier is consistently better than others. Additionally, due to the fact that no other models significantly outperformed other models, it also cannot be determined that classifiers significantly improve in performance as the keystrokes included in the datasets increase. Only the kNN and RF classifiers at 5000 keystrokes significantly outperformed its own model type at lesser keystrokes. In the case of the kNN classifier, it only significantly outperformed the classifier using 1000 keystrokes when using Accuracy as a performance metric. Additionally, while the RF classifier at 5000 keystrokes outperformed RF classifiers using less keystrokes, there was no significant difference between the 1000-keystroke and 2000-keystroke datasets for any performance metric. Since this pattern was also not found for any other models, we conclude that there is no evidence for a gradual increase while using increased amounts of keystrokes.

The finding that a RF classifier performs well in high-dimensional and imbalanced data is confirmed by previous studies (Caruana & Niculescu-Mizil, 2006; Caruana, Karampatziakis, & Yessenalina, 2008; Brown & Mues, 2011). Caruana, Karampatziakis, and Yessenalina (2008) compared 10 different machine learning algorithms on 11 different datasets and found that RF was the best performing classifier in high-dimensional data. However, they also found that LR performed quite well with high-dimensional data as well. A possible reason that LR did not do well in our dataset is likely due to multicollinearity in our datasets. Although we took measures to minimize this and limit features to only those which were relevant with use of the GA and LASSO, it is likely that many features used were still correlated with each other. In addition to this, logistic regression cannot handle complex relationships between predictors and outcome (Dreiseitl & Ohno-Machado, 2002; Kuhle et al., 2018); the fact that LR did not do well at classifying PD stage in the current study implies that the relationship between typing behavior and PD severity is likely complex and therefore unsuitable for LR as a classification method. Instead, we should rely on more complex machine learning algorithms that can pick up on complex relationships in the data.

When examining literature concerning the use of kNN classifiers in high-dimensional data, it also helps explain why this model did not tend to do very well at predicting PD stage. In a study by Sušac, Pfeifer, and Šarlija (2014), it was found that kNN did significantly worse than other machine learning methods in datasets with high dimensionality. In addition, in a study comparing ten different kinds of machine learning methods with each other in high-dimensional data, it was also found that while kNN was among one of the best performers while using lesser amounts of features, it did poorly as the number of features increased, likely due to the amount of noise in the data (Caruana, Karampatziakis, & Yessenalina, 2008). While we attempted to minimize the amount of noise in the data with the use of a GA, the poor performance of the kNN classifier implies that there was likely still noise in the high-dimensional data that affected its performance.

In the current study, we expected all performance metrics to increase as we included more keystrokes in the datasets. However, models generally did not significantly improve as more keystrokes were included. A possible reason for this might simply be that 1000 and 2000 keystrokes were too few to reveal anything meaningful concerning PD stage. It could be that if we used more keystrokes than 5000, we would start to see that classifiers perform significantly better than their counterparts at lesser amounts of keystrokes. In the case of Accuracy, although we did not see that models consistently performed significantly better than models at lesser keystrokes, Figure 1 shows there was a general trend of increased Accuracy as more keystrokes are used. This shows that the use of Accuracy alone in machine learning studies could be misleading, especially when using imbalanced datasets. The current study is a good example of a case where multiple metrics are useful when machine learning is used in order to reduce the possibility that a single metric gives a false interpretation of a classifier's performance.

An additional surprising finding of this study is that Levodopa-use could be predicted in participants using PD significantly better than chance. Levodopa is the most common antiparkinsonian medication on the market (Monteiro, Souza-Machado, Valderramas, & Melo, 2012); it relieves symptoms such as bradykinesia, rigidity, and tremor. In this study, participants using Levodopa were removed to avoid them contributing noise to the data. Although patients taking other antiparkinsonian medications such as dopamine agonists and monoamine oxidase inhibitors were still included, we were nonetheless able to predict PD stage. This implies that Levodopa would be a useful factor to include in models predicting PD stage. While the current study shows that you can predict mild and moderate disease stage with keyboard characteristics alone in patients not using Levodopa, real-world applications would need to be able to differentiate between patients who are also using Levodopa. Therefore, future studies could focus on predicting PD stage with Levodopa-use as an additional predictor. Additionally, this finding may warrant further investigation into the effectiveness of symptom-reducing pharmaceuticals used for PD since the current study indicates that Levodopa is perhaps a more effective symptom-reducing drug than dopamine agonists and monoamine oxidase inhibitors; these medications were still taken by participants included in the datasets but did not distract from being able to predict between mild and moderate PD stage. Adams (2017) had also filtered out these participants in his own study, indicating that he found a similar effect on his results. In this study, removing Levodopa-users from the datasets resulted in few participants in the moderate and severe classes, resulting in heavily imbalanced class distributions. Since Levodopa-use increases as PD severity progresses, future studies should keep Levodopa-use as a predicting factor for machine learning models to prevent this; keeping all possible participants would then help reduce the class imbalance.

Although the results of the current study are promising, there were some limitations. First, the study would have benefited from including more participants since it becomes easier to differentiate between classes when there is plenty of training data; the fact that small sample sizes can lead to inadequate learning for machine learning models (Kotsiantis, Kanellopoulos, & Pintelas, 2006) may have been a contributing factor for poor performance in some models in this study. Furthermore, there were very few participants in the moderate group, and too few in the severe group to use it for machine learning, resulting in the use of oversampling in order to accurately predict the moderate class, and the dropping of the control group. The use of oversampling can help increase the number of correctly classified cases in the minority class; however, it also increases the likelihood of overfitting (Kubat & Matwin, 1997) due to the replication of minority class cases in the training set. In turn, this may have contributed to poor-performing classification in some cases.

An additionally limitation in the current study includes that we used subjective data. In order to report the stage of PD for each participant, participants were asked to self-report what their disease severity was and how much they thought PD impacted their daily lives (Adams, 2017). It might have been more useful to use physicians' assessments of patients' motor skills, since there are many potential inherent biases in self-reported data (Hoskin, 2012). This would be a more clear, direct link to how PD affects typing. As an example, an individual with PD might not find that their disease severity is severe if they have a strong support network or help around the house, whereas someone relying on only themselves might find that their PD is severe since it affects the quality of their daily lives much more. The subjectivity of participants' answers might have left a discrepancy between the PD stage at which they thought they were and the PD stage in which they actually were, meaning that their typing characteristics could more closely match their actual PD stage. Additionally, it could be the case that patients exhibit some symptoms more severely than other symptoms. It would be useful to narrow down which symptoms directly affect typing behavior in case symptom severity does not necessarily correlate with disease severity. Any future studies in this area may benefit from physician input or standardized methods to collect symptom and disease stage. On top of this, future studies would benefit from collecting more data about their participants such as profession, education, and comorbidities; this study was limited to include data on only participants’ PD information. Including more information about participants in the future would allow us to identify patterns within subgroups or factors that contribute to model performance.

The results of the current study lend inspiration to many different future research ideas to expand upon this idea. It appears from this study that keystroke information can provide insight into PD stage. However, as previously mentioned, it would be interesting to see if this could be increased with a higher number of participants and with objective data on PD stage. A promising future project could use UPDRS scores instead of participants’ subjective ideas about how their PD affects them. Using UPDRS scores would also allow more comprehensive predictions of PD severity since they run on a detailed, numeric scale. The use of broad classes of PD severity in this study meant that the models could not learn to distinguish small differences between patients. Additionally, it would be beneficial to include more participants and include participants who use Levodopa, so that the models might be more generalizable and applicable in real-world monitoring systems.

Furthermore, it would be very interesting to examine more computer behavior in PD patients. Examining multiple aspects of computer behavior may help in the monitoring of a wider scope of PD symptoms since keystrokes may only reveal information about specific symptoms. It would be beneficial to do further research to answer the question of which symptoms directly impact different aspects of computer behavior. There are many different aspects of computer interaction that might change as PD progresses including keystroke behavior, mouse clicks and cursor movements, time of day which they use their computers (might be tailored around medication schedule), and the pressure with which individuals press their keys and computer mouse devices. Specialized equipment and software would be needed for this kind of study in order to keep track of these computer behaviors each day for individuals. Additionally, the current study examines PD stage at one point in time. It would be interesting to collect data over time to see how individuals progress. For this, we would have to set up a longitudinal study examining different computer interaction behaviors that may be affected by PD such as the aforementioned characteristics. By answering these questions in the future and creating machine learning models that can detect symptom and disease severity, it becomes possible for physicians and patients to make use of more objective monitoring practices.

The results of the current study indicate that computer interaction can be monitored to track disease progression in patients with PD. Specifically, it shows that keystrokes can be used to predict mild and moderate PD stage with moderate accuracy. Although the current study did not find that keystrokes alone would be accurate enough to use as a monitoring tool, it is the first study to examine whether keystroke characteristics can be used for this purpose. Perhaps by incorporating other monitoring programs, it could help to improve the way that PD patients are currently monitored for their disease progression. Standardized typing tests could be developed for patients with PD so that doctors and patients have the ability to monitor patient symptoms and give insight into their disease progression. This is a promising first step in finding an inexpensive solution for monitoring PD progression.

References

Adams, W. R. (2017). High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing. *PloS one, 12*(11), e0188226.

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa, and the true skill statistic (TSS), *Journal of Applied Ecology*, 43, 1223-1232. doi: 10.1111/j.1365-2664.2006.01214.x

Andlin- Sobocki, P., Jönsson, B., Wittchen, H. U., & Olesen, J. (2005). Cost of disorders of the brain in Europe. *European Journal of neurology, 12*, 1-27. doi: 10.1111/j.1468-1331.2005.01202.x

Australian Bureau of Statistics, 3101.0—Australian Demographic Statistics, Dec 2015, Australian Government, 2016, http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0.

Australian Institute of Health and Welfare, "Australian hospital statistics 2009-2010", Health Services Series no. 40. Cat. no. HSE 107, AIHW, Canberra, Australia, 2011.

Bartmann, D., Bakdi, I., & Achatz, M. (2007). On the design of an authentication system based on keystroke dynamics using a predefined input text. *International Journal of Information Security and Privacy (IJISP)*, *1*(2), 1-12. doi: 10.4018/jisp.2007040101

Bayestehtashk, A., Asgari, M., Shafran, I., & McNames, J. (2015). Fully automated assessment of the severity of Parkinson's disease from speech. *Computer speech & language*, *29*(1), 172-185. doi: 10.1016/j.csl.2013.12.001

Ben-David, A. (2008a). About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence, 21*(6), 874-882. doi: 10.1016/j.engappai.2007.09.009

Ben-David, A. (2008b). Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications, 34*(2), 825-832. doi: 10.1016/j.eswa.2006.10.022

Bohingamu Mudiyanselage, S., Watts, J. J., Abimanyi-Ochom, J., Lane, L., Murphy, A. T., Morris, M. E., & Iansek, R. (2017). Cost of living with Parkinson's disease over 12 months in Australia: A prospective cohort study. *Parkinson's Disease, 2017*. doi: 10.1155/2017/5932675

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. doi:10.1023/A:101093340

Breiman, L. 1999. Using adaptive bagging to debias regressions. Technical Report 547, Statistics Dept. UCB.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453. doi: 10.1016/j.eswa.2011.09.033

Brunato, M., Battiti, R., Pruitt, D., & Sartori, E. (2013). Supervised and unsupervised machine learning for the detection, monitoring and management of Parkinson's disease from passive mobile phone data. *Kaggle Competition. Available at: https://kaggle2. blob. core. windows. net/prospectorfiles/1117/958625cf-3514-4e64-b0e7-13ebd3cf9791/kaggle. pdf.*

Cancela, J., Pastorino, M., Arredondo, M. T., Nikita, K. S., Villagra, F., & Pastor, M. A. (2014). Feasibility study of a wearable system based on a wireless body area network for gait assessment in Parkinson's disease patients. *Sensors*, *14*(3), 4618-4633. doi: 10.3390/s140304618

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). ACM. doi: 10.1145/1143844.1143865

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008, July). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (pp. 96-103). ACM. doi: 10.1145/1390156.1390169

Costa, A., Peppe, A., Brusa, L., Caltagirone, C., Gatto, I., & Carlesimo, G. A. (2008). Levodopa improves time-based prospective memory in Parkinson's disease. *Journal of the International Neuropsychological Society*, *14*(4), 601-610. doi:10.1017/S135561770808082X

Das, R. K., Mukhopadhyay, S., & Bhattacharya, P. (2014). User authentication based on keystroke dynamics. *IETE Journal of Research*, *60*(3), 229-239. doi: 10.1080/03772063.2014.914686

Del Hoyo, L. V., Isabel, M. P. M., & Vega, F. J. M. (2011). Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data. *European Journal of Forest Research*, *130*(6), 983-996. doi: 10.1007/s10342-011-0488-2

Dorsey, E., Constantinescu, R., Thompson, J. P., Biglan, K. M., Holloway, R. G., Kieburtz, K., Marshall, F. J., Ravina, B. M., Schifitto, G, Siderowf, A, & Tanner, C. M. (2007). Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, *68*(5), 384-386. doi: 10.1212/01.wnl.0000247740.47667.03

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics, 35*(5-6), 352-359. doi: 10.1016/S1532-0464(03)00034-0

Drummond, C., & Holte, R. C. (2003, August). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II* (Vol. 11, pp. 1-8). Washington DC: Citeseer.

Duan R-Y, Kong X-Q, Huang M-Y, Fan W-Y, Wang Z-G (2014) The Predictive Performance and Stability of Six Species Distribution Models. PLoS ONE 9(11): e112764. doi: 10.1371/journal.pone.0112764

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007). Learning on the border: active learning in imbalanced data classification. *In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 127-136). ACM. doi: 10.1145/1321440.1321461

Espay, A. J., Bonato, P., Nahab, F. B., Maetzler, W., Dean, J. M., Klucken, J., Eskofier., B. M., Merola, A., Horak, F.,Lang, A. E., Reilmann, R., Giuffrida, J., Nieuwboer, A., Home, M., Little, M. A., Litvan, I., Simuni, T., Dorsey, E. R., Burack, M. A., Kubota, K., Kamondi, A., Godinho, C., Daneault, J., Mitsi, G., Krinke, L., Hausdorff, J. M., Bloem, B. R., Papapetropoulos, S. (2016). Technology in Parkinson's disease: Challenges and opportunities. *Movement Disorders*, *31*(9), 1272-1282. doi: 10.1002/mds.26642

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27-38. doi: 10.1016/j.patrec.2008.08.010

Giot, R., & Rosenberger, C. (2012). A new soft biometric approach for keystroke dynamics based on gender recognition. *International Journal of Information Technology and Management*, *11*(1-2), 35-49. doi: 10.1504/IJITM.2012.044062

Glander, S. (2017, January 15). Feature selection in machine learning (breast cancer datasets). Retrieved from https://shiring.github.io/machine_learning/2017/01/15/rfe_ga_post

Goldberg, D. E., 1989, Genetic Algorithms in Search, Optimization, and Machine Learning (USA: Addison-Wesley Longman, Reading, Massachusetts).

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning*, *3*(2), 95-99. doi: 10.1023/A:102260201

Goodman, J., Syme, A., & Eisma, R. (2003, September). Older adults' use of computers: A survey. In *Proceedings of HCI* (Vol. 2, pp. 25-38).

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E. Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., Gannon, B., Jones, D. H., Jennum, P., Jordanova, A., Jönsson, L., Karampampa, K., Knapp, M., Kobelt, G., & Olesen, J. (2011). Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, *21*(10), 718-779. doi: 10.1016/j.euroneuro.2011.08.008

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, *77*(1), 103-123. doi: 10.1007/s10994-009-5119-5

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186. doi: 10.1023/A:1010920819831

Hoehn, M. M., & Yahr, M. D. (2001). Parkinsonism: Onset, progression and mortality. *Neurology*.

Hoskin, R. (2012). The dangers of self-report. *Science Brainwaves*.

Hosmer, D.W., Lemeshow, S., 1989.Applied Logistic Regression.Wiley, NewYork.

Huang, J., Lu, J., & Ling, C. X. (2003, November). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. In *Null* (p. 553). IEEE. doi: 10.1109/ICDM.2003.1250975

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry*, *79*(4), 368-376. doi: 10.1136/jnnp.2007.131045 PMID: 18344392

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data Recommendations for the Use of Performance Metrics. *International Conference on Affective Computing and Intelligent Interaction and Workshops: [proceedings]. ACII (Conference)*, *2013*, 245–251. doi: 10.1109/ACII.2013.47

Kang, S. J., Choi, J. H., Kim, Y. J., Ma, H. I., & Lee, U. (2015). Development of an acquisition and visualization of forearm tremors and pronation/supination motor activities in a smartphone based environment for an early diagnosis of Parkinson's disease. *Advanced Science and Technology Letters*, *116*, 209-12. doi: 10.14257/astl.2015.116.42

Kaymak, U., Ben-David, A., & Potharst, R. (2012). The AUK: A simple alternative to the AUC. *Engineering Applications of Artificial Intelligence*, *25*(5), 1082-1089.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 25-36. doi: 10.1016/j.engappai.2012.02.012

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*, 3-24.

Kowal, S. L., Dall, T. M., Chakrabarti, R., Storm, M. V., & Jain, A. (2013). The current and projected economic burden of Parkinson's disease in the United States. *Movement Disorders*, *28*(3), 311-318. doi: 10.1002/mds.25292

Kubota, K. J., Chen, J. A., & Little, M. A. (2016). Machine learning for large- scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. *Movement disorders*, *31*(9), 1314-1326. doi: 10.1002/mds.26693

Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, pp. 179-186).

Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A., Joseph, K. S., Allen, V. M. (2018). Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: A retrospective cohort study, *BMC Pregnancy and Childbirth*, *18*(333), 1-9. doi: 10.1186/s12884-018-1971-2

Kuhn, Max. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). caret: Classification and Regression Training. R package version 6.0-78. https://CRAN.R-project.org/package=caret

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. doi: 10.2307/2529310

Lantz, C. A., & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, *49*(4), 431-434. doi: 10.1016/0895-4356(95)00571-4

Ledger, D., & McCaffrey, D. (2014). Inside wearables: How the science of human behavior change offers the secret to long-term engagement. *Endeavour Partners*, *200*(93), 1.

de Lima, A. L. S., Hahn, T., de Vries, N. M., Cohen, E., Bataille, L., Little, M. A., Baldus, H., Bloem, B. R., & Faber, M. J. (2016). Large-scale wearable sensor deployment in Parkinson's patients: the Parkinson@ Home Study Protocol. *JMIR Research Protocols*, *5*(3). doi: 10.2196/resprot.5990

Little, M. A., Costello, D. A., & Harries, M. L. (2011). Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures. *Journal of Voice*, *25*(1), 21-31. doi: 10.1016/j.jvoice.2009.04.004

Mechref, Y. S., Novotny, M. V., Kyselova, Z., & Kang, P. (2008). *U.S. Patent Application No. 12/127,366*.

Men, Y. (2013, March 24). *The Parkinson's voice initiative: Early diagnosis for Parkinson's disease through speech recognition*. Retrieved from https://web.stanford.edu/group/sjph/cgi-bin/sjphsite/the-parkinsons-voice-initiative-early-diagnosis-for-parkinsons-disease-through-speech-recognition/

Monteiro, L., Souza-Machado, A., Valderramas, S., & Melo, A. (2012). The effect of levodopa on pulmonary function in Parkinson's disease: a systematic review and meta-analysis. *Clinical therapeutics*, *34*(5), 1049-1055. doi: 10.1016/j.clinthera.2012.03.00

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. (2003). The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Movement Disorders*, *18*(7), 738-750. doi: 10.1002/mds.10473

The National Health and Medical Research Council (Australia).National Statement on Ethical Conduct in Human Research. The National Health and Medical Research Council, The Australian Research Council, The Australian Vice-Chancellors' Committee; 2015.

Neto, E. C., Bot, B. M., Perumal, T., Omberg, L., Guinney, J., Kellen, M., Klein, A., Friend, S. H., & Trister, A. D. (2016). Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iPhone Sensor data. In *Biocomputing 2016: Proceedings of the Pacific Symposium* (pp. 273-284).

Obuchowski, N. A., Lieber, M. L., & Wians, F. H. (2004). ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clinical chemistry*, *50*(7), 1118-1125. Powers, D. M. (2012, April). The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 345-355). Association for Computational Linguistics. doi: 10.1373/clinchem.2004.031823

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., Jönsson, B., CDBE2010 Study Group, & European Brain Council. (2012). The economic cost of brain disorders in Europe. *European Journal of Neurology*, *19*(1), 155-162. doi: 10.1111/j.1468-1331.2011.03590.x

Patel, S., Lorincz, K., Hughes, R., Huggins, N., Growdon, J., Standaert, D., Akay, M., Dy, J., Welsh, M., & Bonato, P. (2009). Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, *13*(6), 864-873. doi: 10.1109/TITB.2009.2033471

Parkinson's Australia. Parkinson's—Description, Incidence and Theories of Causation [Internet]. 2013. Available from: http://www.parkinsons.org.au/information_sheets

Python Software Foundation. Python Language Reference, version 2.7. Available at https://www.python.org

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reed III, J. F. (2000). Homogeneity of kappa statistics in multiple samples. *Computer methods and programs in biomedicine*, *63*(1), 43-46. doi: 10.1016/S0169-2607(00)00074-2

Roth, G., & Levine, M. D. (1994). Geometric primitive extraction using a genetic algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(9), 901-905. doi: 10.1109/34.310686

Ryan, T. P. (2008). *Modern regression methods* (Vol. 655). John Wiley & Sons.

Siedlecki, W., & Sklansky, J. (1993). A note on genetic algorithms for large-scale feature selection. In *Handbook of Pattern Recognition and Computer Vision* (pp. 88-107).

Sun, Z., Bebis, G., Yuan, X., & Louis, S. J. (2002). Genetic feature subset selection for gender classification: A comparison study. In *Applications of Computer Vision, 2002. (WACV 2002).*

Proceedings. Sixth IEEE Workshop on (pp. 165-170). IEEE. doi: 10.1109/ACV.2002.1182176

Zekić-Sušac, M., Pfeifer, S., & Šarlija, N. (2014). A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem. *Business Systems Research Journal*, *5*(3), 82-96. doi: 10.2478/bsrj-2014-0021

Szöllősi, D., Dénes, D. L., Firtha, F., Kovács, Z., & Fekete, A. (2012). Comparison of six multiclass classifiers by the use of different classification performance indicators. *Journal of Chemometrics*, *26*(3-4), 76-84. doi: 10.1002/cem.2432

Teh, P. S., Teoh, A. B. J., Tee, C., & Ong, T. S. (2011). A multiple layer fusion approach on keystroke dynamics. *Pattern Analysis and Applications*, *14*(1), 23-36.

von Campenhausen, S., Winter, Y., e Silva, A. R., Sampaio, C., Ruzicka, E., Barone, P., Poewe, W., Guekht, A., Mateus, C., Pfeiffer, K., Berger, K, Skoupa, J., Bötzel, K., Geiger-Gritsch, S., Siebert, U., Balzer-Geldsetzer, M., Oertel, W., Dodel, R., & Reese, J. P. (2011). Costs of illness and care in Parkinson's disease: an evaluation in six countries. *European Neuropsychopharmacology*, *21*(2), 180-191.

Wang, S., & Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(4), 1119-1130. doi: 10.1109/TSMCB.2012.2187280

Willis, A. W., Schootman, M., Evanoff, B. A., Perlmutter, J. S., & Racette, B. A. (2011). Neurologist care in Parkinson disease: A utilization, outcomes, and survival study. *Neurology*, *77*(9), 851-857. doi: 10.1212/WNL.0b013e31822c9123

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection* (pp. 117-136). Springer, Boston, MA. doi: 10.1007/978-1-4615-5725-8_8

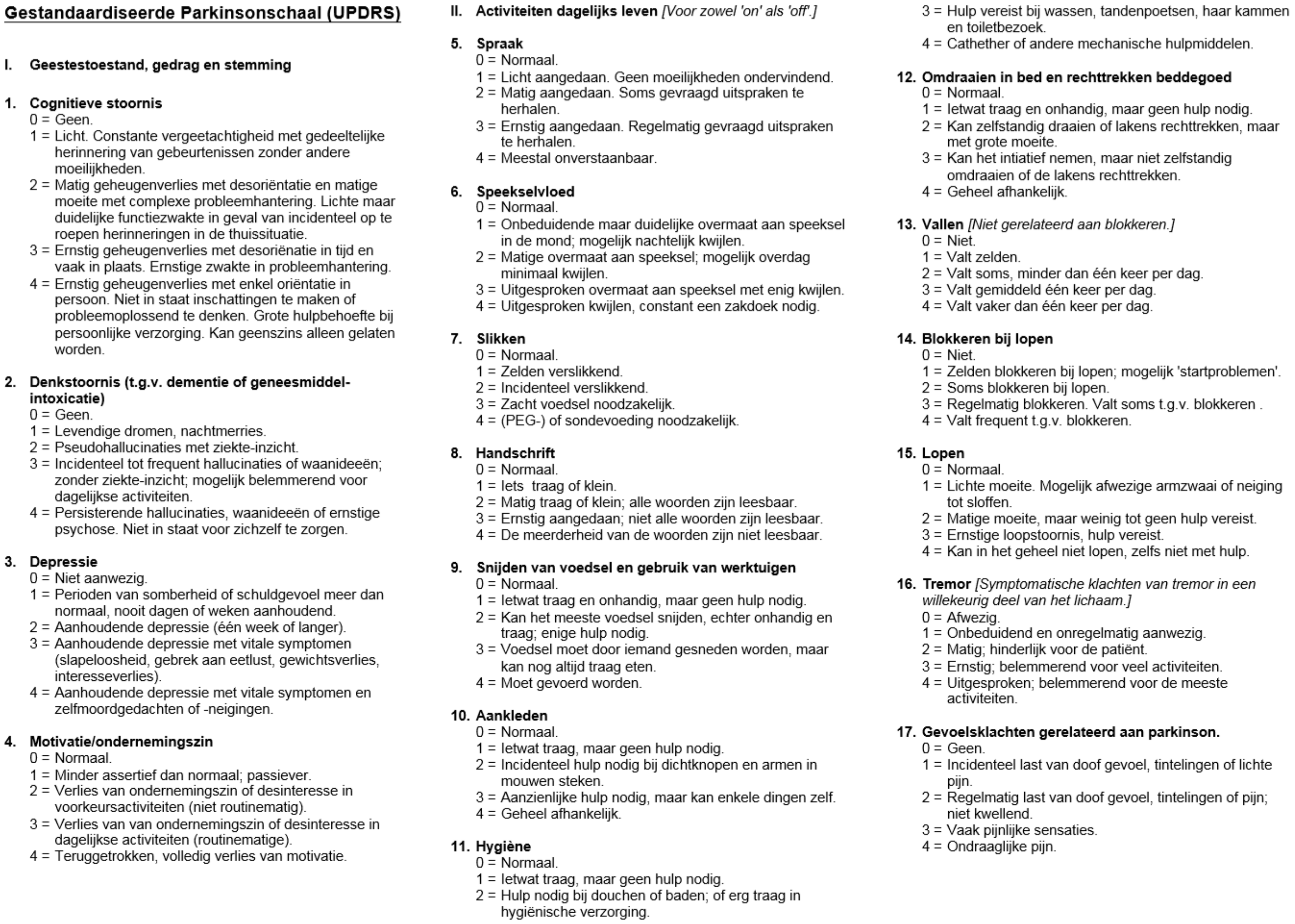
Zamalloa, M., Rodriguez-Fuentes, L. J., Peñagarikano, M., Bordel, G., & Uribe, J. P. (2008, July). Comparing genetic algorithms to principal component analysis and linear discriminant analysis in reducing feature dimensionality for speaker recognition. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation* (pp. 1153-1154). ACM.

Zhan, A., Little, M. A., Harris, D. A., Abiola, S. O., Dorsey, E., Saria, S., & Terzis, A. (2016). High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection. *arXiv preprint arXiv:1601.00960.*

Zhang, L., Deng, J., Pan, Q., Zhan, Y., Fan, J. B., Zhang, K., & Zhang, Z. (2016). Targeted methylation sequencing reveals dysregulated Wnt signaling in Parkinson disease. *Journal of Genetics and Genomics*, *43*(10), 587-592. doi: 10.1016/j.jgg.2016.05.002

APPENDIX A

Dutch Scoring Sheet of UPDRS



Gestandaardiseerde Parkinsonschaal (UPDRS)

I. Geestestoestand, gedrag en stemming

1. Cognitieve stoornis

0 = Geen.

1 = Licht. Constante vergeetachtigheid met gedeeltelijke herinnering van gebeurtenissen zonder andere moeilijkheden.

2 = Matig geheugenverlies met desoriëntatie en matige moeite met complexe probleemhantering. Lichte maar duidelijke functiezwakte in geval van incidenteel op te roepen herinneringen in de thuissituatie.

3 = Ernstig geheugenverlies met desoriënatie in tijd en vaak in plaats. Ernstige zwakte in probleemhantering.

4 = Ernstig geheugenverlies met enkel oriëntatie in persoon. Niet in staat inschattingen te maken of probleemoplossend te denken. Grote hulpbehoefte bij persoonlijke verzorging. Kan geenszins alleen gelaten worden.

2. Denkstoornis (t.g.v. dementie of geneesmiddel-intoxicatie)

0 = Geen.

1 = Levendige dromen, nachtmerries.

2 = Pseudohallucinaties met ziekte-inzicht.

3 = Incidenteel tot frequent hallucinaties of waanideeën; zonder ziekte-inzicht; mogelijk belemmerend voor dagelijkse activiteiten.

4 = Persisterende hallucinaties, waanideeën of ernstige psychose. Niet in staat voor zichzelf te zorgen.

3. Depressie

0 = Niet aanwezig.

1 = Perioden van somberheid of schuldgevoel meer dan normaal, nooit dagen of weken aanhoudend.

2 = Aanhoudende depressie (één week of langer).

3 = Aanhoudende depressie met vitale symptomen (slapeloosheid, gebrek aan eetlust, gewichtsverlies, interesseverlies).

4 = Aanhoudende depressie met vitale symptomen en zelfmoordgedachten of -neigingen.

4. Motivatie/ondernemingszin

0 = Normaal.

1 = Minder assertief dan normaal; passiever.

2 = Verlies van ondernemingszin of desinteresse in voorkeursactiviteiten (niet routinematig).

3 = Verlies van van ondernemingszin of desinteresse in dagelijkse activiteiten (routinematige).

4 = Teruggetrokken, volledig verlies van motivatie.

II. Activiteiten dagelijks leven *[Voor zowel 'on' als 'off'.]*

5. Spraak

0 = Normaal.

1 = Licht aangedaan. Geen moeilijkheden ondervindend.

2 = Matig aangedaan. Soms gevraagd uitspraken te herhalen.

3 = Ernstig aangedaan. Regelmatig gevraagd uitspraken te herhalen.

4 = Meestal onverstaanbaar.

6. Speekselvloed

0 = Normaal.

1 = Onbeduidende maar duidelijke overmaat aan speeksel in de mond; mogelijk nachtelijk kwijlen.

2 = Matige overmaat aan speeksel; mogelijk overdag minimaal kwijlen.

3 = Uitgesproken overmaat aan speeksel met enig kwijlen.

4 = Uitgesproken kwijlen, constant een zakdoek nodig.

7. Slikken

0 = Normaal.

1 = Zelden verslikkend.

2 = Incidenteel verslikkend.

3 = Zacht voedsel noodzakelijk.

4 = (PEG-) of sondevoeding noodzakelijk.

8. Handschrift

0 = Normaal.

1 = Iets traag of klein.

2 = Matig traag of klein; alle woorden zijn leesbaar.

3 = Ernstig aangedaan; niet alle woorden zijn leesbaar.

4 = De meerderheid van de woorden zijn niet leesbaar.

9. Snijden van voedsel en gebruik van werktuigen

0 = Normaal.

1 = Ietwat traag en onhandig, maar geen hulp nodig.

2 = Kan het meeste voedsel snijden, echter onhandig en traag; enige hulp nodig.

3 = Voedsel moet door iemand gesneden worden, maar kan nog altijd traag eten.

4 = Moet gevoerd worden.

10. Aankleden

0 = Normaal.

1 = Ietwat traag, maar geen hulp nodig.

2 = Incidenteel hulp nodig bij dichtknopen en armen in mouwen steken.

3 = Aanzienlijke hulp nodig, maar kan enkele dingen zelf.

4 = Geheel afhankelijk.

11. Hygiène

0 = Normaal.

1 = Ietwat traag, maar geen hulp nodig.

2 = Hulp nodig bij douchen of baden; of erg traag in hygiënische verzorging.

3 = Hulp vereist bij wassen, tandenpoetsen, haar kammen en toiletbezoek.

4 = Cathether of andere mechanische hulpmiddelen.

12. Omdraaien in bed en rechttrekken beddegoed

0 = Normaal.

1 = Ietwat traag en onhandig, maar geen hulp nodig.

2 = Kan zelfstandig draaien of lakens rechttrekken, maar met grote moeite.

3 = Kan het intiatief nemen, maar niet zelfstandig omdraaien of de lakens rechttrekken.

4 = Geheel afhankelijk.

13. Vallen *[Niet gerelateerd aan blokkeren.]*

0 = Niet.

1 = Valt zelden.

2 = Valt soms, minder dan één keer per dag.

3 = Valt gemiddeld één keer per dag.

4 = Valt vaker dan één keer per dag.

14. Blokkeren bij lopen

0 = Niet.

1 = Zelden blokkeren bij lopen; mogelijk 'startproblemen'.

2 = Soms blokkeren bij lopen.

3 = Regelmatig blokkeren. Valt soms t.g.v. blokkeren .

4 = Valt frequent t.g.v. blokkeren.

15. Lopen

0 = Normaal.

1 = Lichte moeite. Mogelijk afwezige armzwaai of neiging tot sloffen.

2 = Matige moeite, maar weinig tot geen hulp vereist.

3 = Ernstige loopstoornis, hulp vereist.

4 = Kan in het geheel niet lopen, zelfs niet met hulp.

16. Tremor *[Symptomatische klachten van tremor in een willekeurig deel van het lichaam.]*

0 = Afwezig.

1 = Onbeduidend en onregelmatig aanwezig.

2 = Matig; hinderlijk voor de patiënt.

3 = Ernstig; belemmerend voor veel activiteiten.

4 = Uitgesproken; belemmerend voor de meeste activiteiten.

17. Gevoelsklachten gerelateerd aan parkinson.

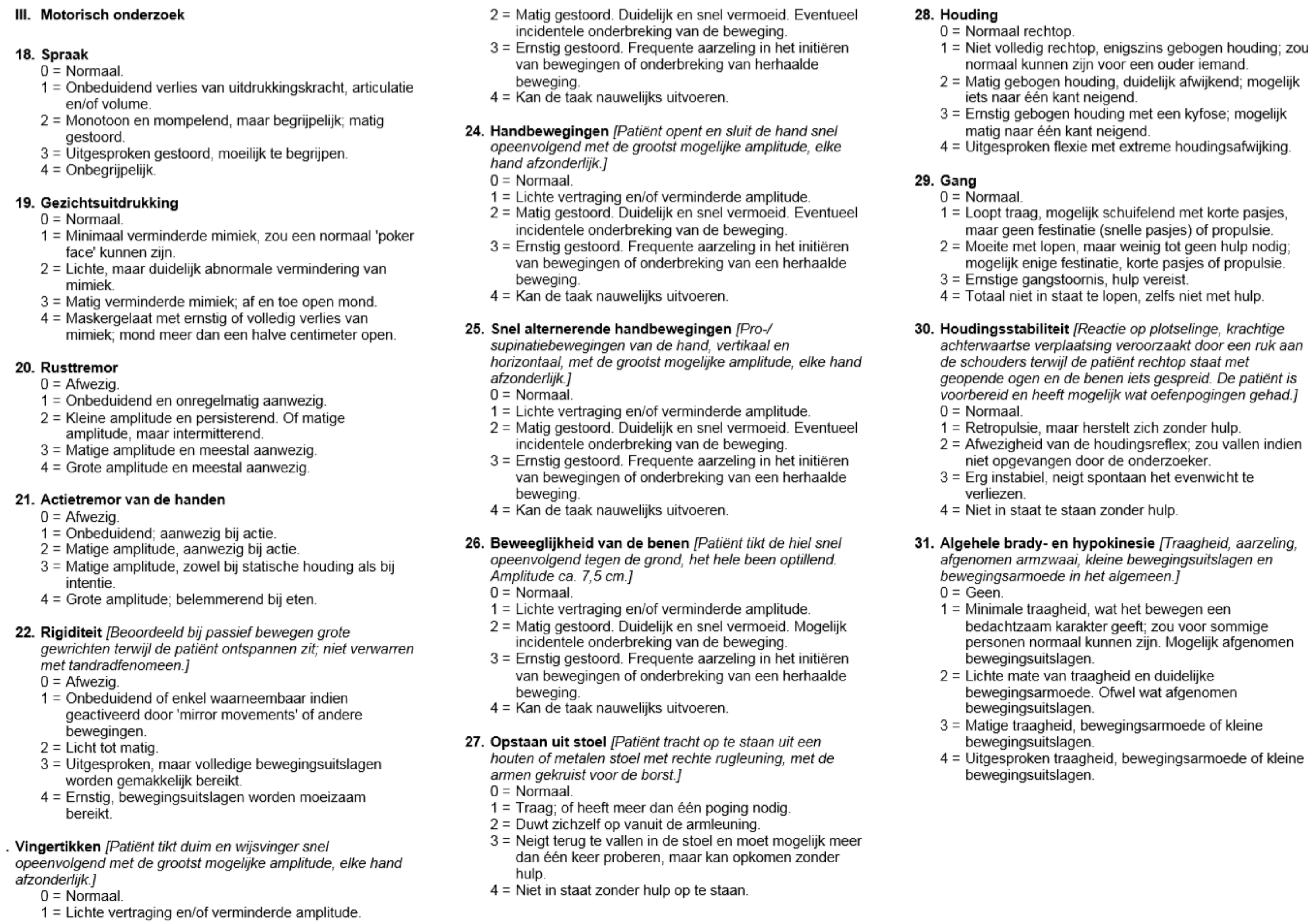
0 = Geen.

1 = Incidenteel last van doof gevoel, tintelingen of lichte pijn.

2 = Regelmatig last van doof gevoel, tintelingen of pijn; niet kwellend.

3 = Vaak pijnlijke sensaties.

4 = Ondraaglijke pijn.



III. Motorisch onderzoek

18. Spraak
0 = Normaal.
1 = Onbeduidend verlies van uitdrukkingskracht, articulatie en/of volume.
2 = Monotoon en mompelend, maar begrijpelijk; matig gestoord.
3 = Uitgesproken gestoord, moeilijk te begrijpen.
4 = Onbegrijpelijk.

19. Gezichtsuitdrukking
0 = Normaal.
1 = Minimaal verminderde mimiek, zou een normaal 'poker face' kunnen zijn.
2 = Lichte, maar duidelijk abnormale vermindering van mimiek.
3 = Matig verminderde mimiek; af en toe open mond.
4 = Maskergelaat met ernstig of volledig verlies van mimiek; mond meer dan een halve centimeter open.

20. Rusttremor
0 = Afwezig.
1 = Onbeduidend en onregelmatig aanwezig.
2 = Kleine amplitude en persisterend. Of matige amplitude, maar intermitterend.
3 = Matige amplitude en meestal aanwezig.
4 = Grote amplitude en meestal aanwezig.

21. Actietremor van de handen
0 = Afwezig.
1 = Onbeduidend; aanwezig bij actie.
2 = Matige amplitude, aanwezig bij actie.
3 = Matige amplitude, zowel bij statische houding als bij intentie.
4 = Grote amplitude; belemmerend bij eten.

22. Rigiditeit *[Beoordeeld bij passief bewegen grote gewrichten terwijl de patiënt ontspannen zit; niet verwarren met tandradfenomeen.]*
0 = Afwezig.
1 = Onbeduidend of enkel waarneembaar indien geactiveerd door 'mirror movements' of andere bewegingen.
2 = Licht tot matig.
3 = Uitgesproken, maar volledige bewegingsuitslagen worden gemakkelijk bereikt.
4 = Ernstig, bewegingsuitslagen worden moeizaam bereikt.

. Vingertikken *[Patiënt tikt duim en wijsvinger snel opeenvolgend met de grootst mogelijke amplitude, elke hand afzonderlijk.]*
0 = Normaal.
1 = Lichte vertraging en/of verminderde amplitude.
2 = Matig gestoord. Duidelijk en snel vermoeid. Eventueel incidentele onderbreking van de beweging.
3 = Ernstig gestoord. Frequente aarzeling in het initiëren van bewegingen of onderbreking van herhaalde beweging.
4 = Kan de taak nauwelijks uitvoeren.

24. Handbewegingen *[Patiënt opent en sluit de hand snel opeenvolgend met de grootst mogelijke amplitude, elke hand afzonderlijk.]*
0 = Normaal.
1 = Lichte vertraging en/of verminderde amplitude.
2 = Matig gestoord. Duidelijk en snel vermoeid. Eventueel incidentele onderbreking van de beweging.
3 = Ernstig gestoord. Frequente aarzeling in het initiëren van bewegingen of onderbreking van een herhaalde beweging.
4 = Kan de taak nauwelijks uitvoeren.

25. Snel alternerende handbewegingen *[Pro-/ supinatiebewegingen van de hand, vertikaal en horizontaal, met de grootst mogelijke amplitude, elke hand afzonderlijk.]*
0 = Normaal.
1 = Lichte vertraging en/of verminderde amplitude.
2 = Matig gestoord. Duidelijk en snel vermoeid. Eventueel incidentele onderbreking van de beweging.
3 = Ernstig gestoord. Frequente aarzeling in het initiëren van bewegingen of onderbreking van een herhaalde beweging.
4 = Kan de taak nauwelijks uitvoeren.

26. Beweeglijkheid van de benen *[Patiënt tikt de hiel snel opeenvolgend tegen de grond, het hele been optillend. Amplitude ca. 7,5 cm.]*
0 = Normaal.
1 = Lichte vertraging en/of verminderde amplitude.
2 = Matig gestoord. Duidelijk en snel vermoeid. Mogelijk incidentele onderbreking van de beweging.
3 = Ernstig gestoord. Frequente aarzeling in het initiëren van bewegingen of onderbreking van een herhaalde beweging.
4 = Kan de taak nauwelijks uitvoeren.

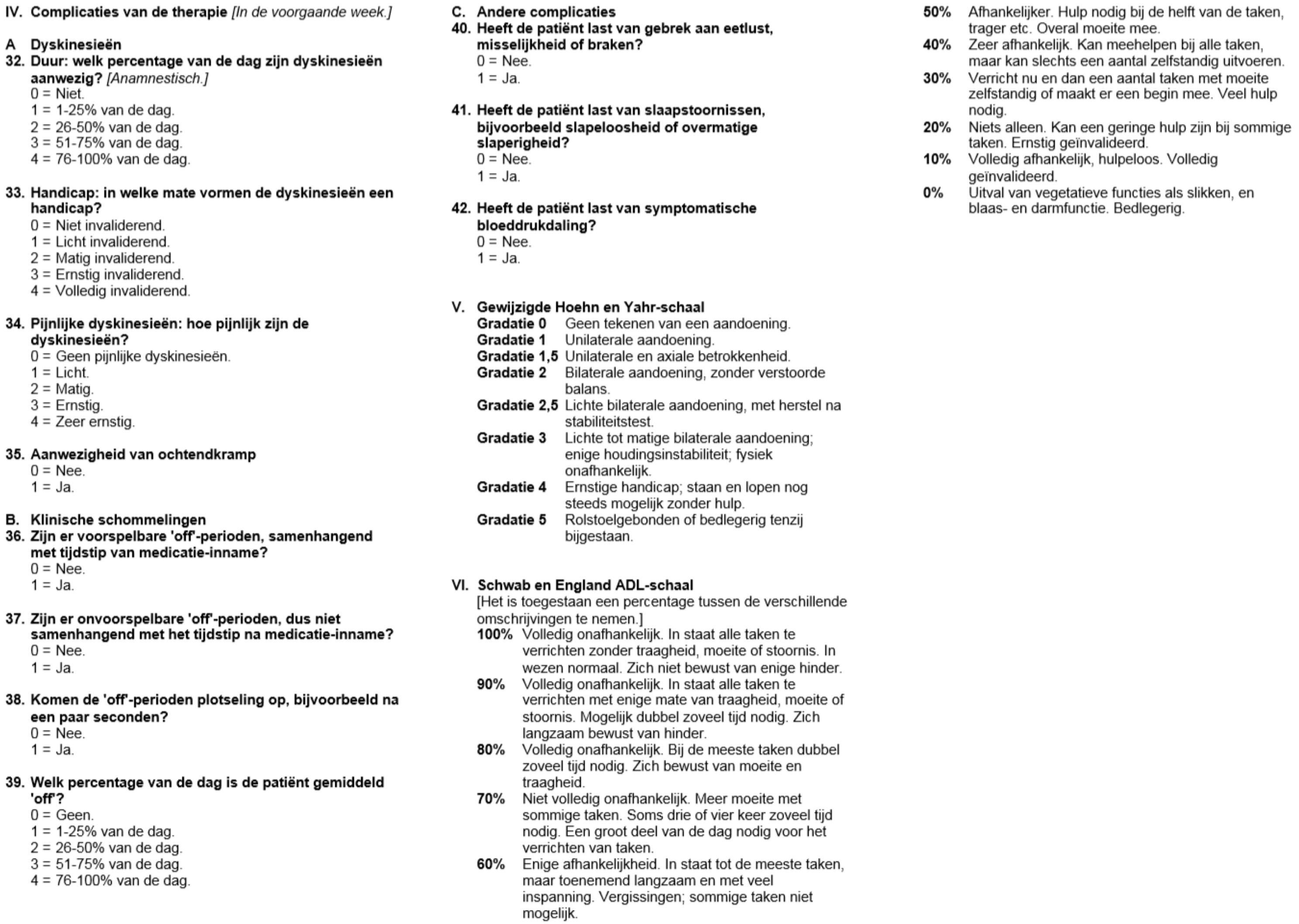
27. Opstaan uit stoel *[Patiënt tracht op te staan uit een houten of metalen stoel met rechte rugleuning, met de armen gekruist voor de borst.]*
0 = Normaal.
1 = Traag; of heeft meer dan één poging nodig.
2 = Duwt zichzelf op vanuit de armleuning.
3 = Neigt terug te vallen in de stoel en moet mogelijk meer dan één keer proberen, maar kan opkomen zonder hulp.
4 = Niet in staat zonder hulp op te staan.

28. Houding
0 = Normaal rechtop.
1 = Niet volledig rechtop, enigszins gebogen houding; zou normaal kunnen zijn voor een ouder iemand.
2 = Matig gebogen houding, duidelijk afwijkend; mogelijk iets naar één kant neigend.
3 = Ernstig gebogen houding met een kyfose; mogelijk matig naar één kant neigend.
4 = Uitgesproken flexie met extreme houdingsafwijking.

29. Gang
0 = Normaal.
1 = Loopt traag, mogelijk schuifelend met korte pasjes, maar geen festinatie (snelle pasjes) of propulsie.
2 = Moeite met lopen, maar weinig tot geen hulp nodig; mogelijk enige festinatie, korte pasjes of propulsie.
3 = Ernstige gangstoornis, hulp vereist.
4 = Totaal niet in staat te lopen, zelfs niet met hulp.

30. Houdingsstabiliteit *[Reactie op plotselinge, krachtige achterwaartse verplaatsing veroorzaakt door een ruk aan de schouders terwijl de patiënt rechtop staat met geopende ogen en de benen iets gespreid. De patiënt is voorbereid en heeft mogelijk wat oefenpogingen gehad.]*
0 = Normaal.
1 = Retropulsie, maar herstelt zich zonder hulp.
2 = Afwezigheid van de houdingsreflex; zou vallen indien niet opgevangen door de onderzoeker.
3 = Erg instabiel, neigt spontaan het evenwicht te verliezen.
4 = Niet in staat te staan zonder hulp.

31. Algehele brady- en hypokinesie *[Traagheid, aarzeling, afgenomen armzwaai, kleine bewegingsuitslagen en bewegingsarmoede in het algemeen.]*
0 = Geen.
1 = Minimale traagheid, wat het bewegen een bedachtzaam karakter geeft; zou voor sommige personen normaal kunnen zijn. Mogelijk afgenomen bewegingsuitslagen.
2 = Lichte mate van traagheid en duidelijke bewegingsarmoede. Ofwel wat afgenomen bewegingsuitslagen.
3 = Matige traagheid, bewegingsarmoede of kleine bewegingsuitslagen.
4 = Uitgesproken traagheid, bewegingsarmoede of kleine bewegingsuitslagen.



IV. Complicaties van de therapie *[In de voorgaande week.]*

A Dyskinesieën

32. Duur: welk percentage van de dag zijn dyskinesieën aanwezig? *[Anamnestisch.]*
0 = Niet.
1 = 1-25% van de dag.
2 = 26-50% van de dag.
3 = 51-75% van de dag.
4 = 76-100% van de dag.

33. Handicap: in welke mate vormen de dyskinesieën een handicap?
0 = Niet invaliderend.
1 = Licht invaliderend.
2 = Matig invaliderend.
3 = Ernstig invaliderend.
4 = Volledig invaliderend.

34. Pijnlijke dyskinesieën: hoe pijnlijk zijn de dyskinesieën?
0 = Geen pijnlijke dyskinesieën.
1 = Licht.
2 = Matig.
3 = Ernstig.
4 = Zeer ernstig.

35. Aanwezigheid van ochtendkramp
0 = Nee.
1 = Ja.

B. Klinische schommelingen

36. Zijn er voorspelbare 'off'-perioden, samenhangend met tijdstip van medicatie-inname?
0 = Nee.
1 = Ja.

37. Zijn er onvoorspelbare 'off'-perioden, dus niet samenhangend met het tijdstip na medicatie-inname?
0 = Nee.
1 = Ja.

38. Komen de 'off'-perioden plotseling op, bijvoorbeeld na een paar seconden?
0 = Nee.
1 = Ja.

39. Welk percentage van de dag is de patiënt gemiddeld 'off'?
0 = Geen.
1 = 1-25% van de dag.
2 = 26-50% van de dag.
3 = 51-75% van de dag.
4 = 76-100% van de dag.

C. Andere complicaties

40. Heeft de patiënt last van gebrek aan eetlust, misselijkheid of braken?
0 = Nee.
1 = Ja.

41. Heeft de patiënt last van slaapstoornissen, bijvoorbeeld slapeloosheid of overmatige slaperigheid?
0 = Nee.
1 = Ja.

42. Heeft de patiënt last van symptomatische bloeddrukdaling?
0 = Nee.
1 = Ja.

V. Gewijzigde Hoehn en Yahr-schaal

Gradatie 0 Geen tekenen van een aandoening.
Gradatie 1 Unilaterale aandoening.
Gradatie 1,5 Unilaterale en axiale betrokkenheid.
Gradatie 2 Bilaterale aandoening, zonder verstoorde balans.
Gradatie 2,5 Lichte bilaterale aandoening, met herstel na stabiliteitstest.
Gradatie 3 Lichte tot matige bilaterale aandoening; enige houdingsinstabiliteit; fysiek onafhankelijk.
Gradatie 4 Ernstige handicap; staan en lopen nog steeds mogelijk zonder hulp.
Gradatie 5 Rolstoelgebonden of bedlegerig tenzij bijgestaan.

VI. Schwab en England ADL-schaal

[Het is toegestaan een percentage tussen de verschillende omschrijvingen te nemen.]

100% Volledig onafhankelijk. In staat alle taken te verrichten zonder traagheid, moeite of stoornis. In wezen normaal. Zich niet bewust van enige hinder.
90% Volledig onafhankelijk. In staat alle taken te verrichten met enige mate van traagheid, moeite of stoornis. Mogelijk dubbel zoveel tijd nodig. Zich langzaam bewust van hinder.
80% Volledig onafhankelijk. Bij de meeste taken dubbel zoveel tijd nodig. Zich bewust van moeite en traagheid.
70% Niet volledig onafhankelijk. Meer moeite met sommige taken. Soms drie of vier keer zoveel tijd nodig. Een groot deel van de dag nodig voor het verrichten van taken.
60% Enige afhankelijkheid. In staat tot de meeste taken, maar toenemend langzaam en met veel inspanning. Vergissingen; sommige taken niet mogelijk.
50% Afhankelijker. Hulp nodig bij de helft van de taken, trager etc. Overal moeite mee.
40% Zeer afhankelijk. Kan meehelpen bij alle taken, maar kan slechts een aantal zelfstandig uitvoeren.
30% Verricht nu en dan een aantal taken met moeite zelfstandig of maakt er een begin mee. Veel hulp nodig.
20% Niets alleen. Kan een geringe hulp zijn bij sommige taken. Ernstig geïnvalideerd.
10% Volledig afhankelijk, hulpeloos. Volledig geïnvalideerd.
0% Uitval van vegetatieve functies als slikken, en blaas- en darmfunctie. Bedlegerig.

APPENDIX B

Confusion Matrices for the 1000 Keyboard Dataset Using All Three Classification Groups.

Models were made using 5-fold cross validation, and then tested on the test set. The dataset to create these models consisted of 1000 keystrokes. The results in this table show that the moderate class (minority class) is rarely predicted correctly (only twice in the LR model) while the control and mild conditions are more likely to be correctly predicted.

LR			
Actual	None	Mild	Moderate
Predicted			
None	4	1	1
Mild	7	3	1
Moderate	4	3	2
kNN			
Actual	None	Mild	Moderate
Predicted			
None	6	0	1
Mild	9	7	3
Moderate	0	0	0
RF			
Actual	None	Mild	Moderate
Predicted			
None	8	1	2
Mild	7	6	2
Moderate	0	0	0

APPENDIX C

Confusion Matrices for the 2000 Keyboard Dataset Using All Three Classification Groups

Models were made using 5-fold cross validation, and then tested on the test set. The dataset to create these models consisted of 2000 keystrokes. The results in this table show that the moderate class (minority class) is rarely predicted (only once in the LR model) correctly while the control and mild conditions are generally more likely to be correctly predicted.

LR			
Actual	None	Mild	Moderate
Predicted			
None	3	1	0
Mild	6	3	2
Moderate	4	2	1
kNN			
Actual	None	Mild	Moderate
Predicted			
None	5	0	0
Mild	8	6	3
Moderate	0	0	0
RF			
Actual	None	Mild	Moderate
Predicted			
None	9	2	0
Mild	4	3	3
Moderate	0	1	0

APPENDIX D

Confusion Matrices for the 5000 Keyboard Dataset Using All Three Classification Groups

Models were made using 5-fold cross validation, and then tested on the test set. The dataset to create these models consisted of 5000 keystrokes. The results in this table show that the moderate class (minority class) is never predicted correctly while the control and mild conditions are often correctly predicted.

LR			
Actual	None	Mild	Moderate
Predicted			
None	1	4	2
Mild	5	3	1
Moderate	1	0	0
kNN			
Actual	None	Mild	Moderate
Predicted			
None	4	0	0
Mild	3	7	3
Moderate	0	0	0
RF			
Actual	None	Mild	Moderate
Predicted			
None	5	0	3
Mild	2	7	0
Moderate	0	0	0

APPENDIX E

Confusion Matrices for the 1000 Keyboard Dataset Using Mild and Moderate Classes

Models were made without the use of cross validation, and then tested on the test set. The dataset to create these models consisted of 1000 keystrokes. The results in this table show that the moderate class (minority class) is rarely correctly predicted in all three model types. However, this is still an improvement in comparison to the number of correctly predicted cases of the moderate class when including the control group.

LR		
Actual	Mild	Moderate
Predicted		
Mild	5	4
Moderate	3	2
kNN		
Actual	Mild	Moderate
Predicted		
Mild	7	5
Moderate	1	1
RF		
Actual	Mild	Moderate
Predicted		
Mild	7	5
Moderate	1	1

APPENDIX F

Confusion Matrices for the 2000 Keyboard Dataset Using Mild and Moderate Classes

Models were made without the use of cross validation, and then tested on the test set. The dataset to create these models consisted of 2000 keystrokes. The results in this table show that the moderate class (minority class) is sometimes correctly predicted in the kNN and RF classifiers, but never in the LR classifier. This is an improvement in comparison to the number of correctly predicted cases of the moderate class when including the control group where it was rarely correctly classified.

LR		
Actual	Mild	Moderate
Predicted		
Mild	7	4
Moderate	1	0
kNN		
Actual	Mild	Moderate
Predicted		
Mild	6	2
Moderate	2	2
RF		
Actual	Mild	Moderate
Predicted		
Mild	7	2
Moderate	1	2

APPENDIX G

Confusion Matrices for the 5000 Keyboard Dataset Using Mild and Moderate Classes

Models were made without the use of cross validation, and then tested on the test set. The dataset to create these models consisted of 5000 keystrokes. The results in this table show that the moderate class (minority class) is never correctly predicted using the LR classifier, once using the kNN classifier, and most often using the RF classifier. This is an improvement in comparison to the number of correctly predicted cases of the moderate class when including the control group where it was never correctly classified.

LR		
Actual	Mild	Moderate
Predicted		
Mild	5	3
Moderate	2	0
kNN		
Actual	Mild	Moderate
Predicted		
Mild	6	2
Moderate	1	1
RF		
Actual	Mild	Moderate
Predicted		
Mild	7	1
Moderate	0	2