Linking Data Collected from Mobile Phones with Symptoms Level in Parkinson's Disease: Data Exploration of the mPower Study

Gent Ymeri(✉), Dario Salvi, and Carl Magnus Olsson

Internet of Things and People, Malmö University, Malmö, Sweden
gent.ymeri@mau.se

Abstract. Advancements in technology, such as smartphones and wearable devices, can be used for collecting movement data through embedded sensors. This paper focuses on linking Parkinson's Disease severity with data collected from mobile phones in the mPower study. As reference for symptoms' severity, we use the answers provided to part 2 of the standard MDS-UPDRS scale. As input variables, we use the features computed within mPower from the raw data collected during 4 phone-based activities: walking, rest, voice and finger tapping. The features are filtered in order to remove unreliable datapoints and associated to reference values. After pre-processing, 5 Machine Learning algorithms are applied for predictive analysis. We show that, notwithstanding the noise due to the data being collected in an uncontrolled manner, the regressed symptom levels are moderately to strongly correlated with the actual values (highest Pearson's correlation = 0.6). However, the high difference between the values also implies that the regressed values can not be considered as a substitute for a conventional clinical assessment (lowest mean absolute error = 5.4).

Keywords: mobile health · Parkinson's disease · mPower data

1 Introduction

Parkinson's disease (PD) is a chronic neurodegenerative disease characterised by a complex symptomatology, including impaired motor function, sleep and neuropsychiatric disorders [1]. It is the second most common neurodegenerative disease and affects more than 6 million people worldwide - a number that is expected to double in 20 years [2].

Diagnosing and assessing PD is based on clinical criteria such as the Unified Parkinson Disease Rating Scale (UPDRS). The first version of the scale was established in the 1980s, while the revised MDS-UPDRS was established in 2008 by the Movement Disorder Society (MDS) [3]. Although highly accepted

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 439–449, 2023.
https://doi.org/10.1007/978-3-031-34586-9_29

in clinical practice, these scales are used intermittently, are based on subjective criteria, and can be unreliable [4]. Subsequently, this negatively affects optimal patients' care.

Technological advancements such as smartphones and wearable devices have the potential to gather objective data for assessing disease severity on PD patients [5], thus addressing the subjectiveness of the UPDRS and MDS-UPDRS scales. To validate the usefulness of smartphones for PD treatment, the mPower observational study consists of longitudinal and frequent data collection from 14,684 individuals, both PD patients and healthy participants [6]. The study includes surveys and activities captured by smartphone sensors. Such activities include memory tests, finger tapping, vocalization test and walking test, while surveys include demographic data and other PD rating scales such as a subset of MDS-UPDRS. Previous research, cf. [7–9], has shown that these data can be used to distinguish between subjects with PD and subjects without PD, but little evidence exists that they can be associated to symptoms levels.

This paper describes an extended analysis of the mPower data set. This is done by first providing an overview to ensure that readers have a general understanding of the data set, then moves on to assessing if it is possible to predict the disease severity level based on the partial, self-reported MDS-UPDRS, together with the data collected within the smartphone-based activities.

2 Related Work

Using mobile applications to monitor health state and evaluate cognitive and motor deficits in patients with diseases that affect the central nervous system can be achieved as shown in [10]. For what regards Parkinson's Disease several studies have tried to link data from sensors and smartphones to severity levels. These include, for example, the quantification of dexterity levels of PD patients through finger tapping and spiral drawing tests using a smartphone [11]. After using machine-learning models on a set of 37 spatiotemporal features, the authors could report weak to moderate correlations between smartphone-based scores and ratings of some motor items from Part III of UPDRS. Hand tremor, another common symptom in PD, was assessed using smartphones' accelerometers in [8]. As part of that study, the authors propose an objective hand tremor severity score based on spectral power features of the acceleration signal that is shown to be significantly correlated to the self-assessed tremor score in UPDRS part II. In another study, gait analysis was conducted using an app to collect data in two 20-meter tests with PD patients walking normally and walking while performing a mental task [7]. Results from this study show how gait features such as stride time variability correlate with the UPDRS part III total score.

The mPower dataset, which we focus on in this paper, has been used in previous studies for predicting dopaminergic medication response using sensors data [9] and in the DREAM Challenge [12], where different teams competed to develop the best algorithm to e.g. differentiate between PD cases and controls. More related to the aim of our work, the mPower dataset has been also used to

associate smartphone-based data with in-clinic assessments in a sub-study with 44 participants over a 6-month period [13]. The study shows how the original dataset, where volunteers were recruited online and were not followed up by any clinician, presents several biases, such as age and gender, which are not balanced when classifying PD vs non-PD. Using the controlled 44 patients cohort, authors could develop an "mPower symptom severity score" which they derived from the prediction probability of being affected by PD generated from the data of each of the activities. They showed that task performance, especially finger tapping, is predictive of self-reported PD status and correlated with in-clinic evaluation of disease severity.

While this study shows that smartphones allow remote, objective and personalized assessments of PD patients [13], the work relies on patients recruited in a controlled study. In this paper, we instead exploit the full 14-thousand uncontrolled volunteers dataset to identify links between smartphone data and symptom levels. As ground truth, we specifically use the subset of part 2 of the MDS-UPDRS self-reported questionnaire that volunteers were asked to answer.

3 Methods

3.1 The mPower Dataset

The data from mPower study was collected through Apple smartphones using ResearchKit [6]. Enrolled participants include people diagnosed with and without PD, with the latter participating as control. Patients were asked to perform 7 tasks using the smartphone: 3 surveys and 4 activity tasks. The surveys include a demographics questionnaire for PD assessment, the Parkinson Disease Questionnaire 8 (PDQ8), and a selection of items from the MDS-UPDRS, particularly questions 1.1 to 2.13, which can be self-reported and do not need clinician's observations.

The non-survey tasks consist of 4 activities: Memory activity, Tapping activity, Voice activity, and Walking activity.

1. Memory activity: used to evaluate short-term spatial memory. This is achieved by asking the participant to observe a grid of flowers that is illuminated in a sequence and to replicate the pattern in the same order by touching the flowers on the screen of the phone.
2. Tapping activity: used to measure dexterity and speed of fingers' movement. This is done by asking participants to tap on the screen of the phone with two fingers, alternatively, for 20 s.
3. Voice activity is used to measure sustained phonation by asking participants to vocalize "Aaaah" steadily for about 10 s.
4. Walking activity: used to evaluate the gait and balance of participants. They are asked to walk in a straight line for about 20 steps, turn around, stand still for 30 s and then walk again for 20 steps to get back to the same spot they started. The standing still phase also worked as a balance test.

In our study, we use the mPower data collected for the motor-related activities as input (tapping, voice, rest and walking), more concretely, the features computed and made available in [6], and the subset of part 2 of the MDS-UPDRS questionnaire, which is related to motor symptoms, as disease level reference.

Each question in the MDS-UPDRS allows one answer on a 5 levels scale, where 'Normal', 'Slight', 'Mild', 'Moderate', and 'Severe' are mapped to 0, 1, 2, 3, and 4 respectively. The answers thus allow us to compute a score for each part of the MDS-UPDRS and a compound one for the whole questionnaire.

In order to compare our results with existing literature such as [1], we strived to predict the full rating of part 2 of MDS-UPDRS. That part of the scale consists of 13 questions with a total score of 0 to 52. However, in mPower only 10 questions are provided (missing questions are 2.2, 2.3 and 2.11), thus reducing the maximum score to 40. As a result, we normalized the scoring by summing all the questions' scores, dividing this score by 40 and multiplying it by 52. The formula looks as follows:

$$Normalized\_Score = \frac{\sum All\_Question\_Scores}{40} * 52 \tag{1}$$

.

3.2 Descriptive Statistics of the Data

The features computed in [6] for the 3 motor-related activities are separated into 4 subsets, because the walking activity is further split into features related to the walking phase of the activity and features related to the rest/balance phase. The number of subjects (including both PD and healthy) differs depending on the different activity data: finger tapping, walking, rest and voice (see Table 1).

Table 1. Total number of unique participants and number of tests for each activity.

	Finger tapping	Walking	Rest	Voice
Number of subjects	8003	3070	3100	5810
Number of tests	78880	34679	35407	64391

As visible in Fig. 1, the dataset is highly skewed, with few participants having performed a high number of tests and most participants having contributed with a few tests.

The number of unique participants for self-reported answers to motion-related MDS-UPDRS questions is 2024. Whereas the total number of answered questionnaires is 2305. Skewness can also be observed for these answers, with most participants (1951) having answered only once (see Fig. 3). In addition, the distribution of the score for motor symptoms computed as in Eq. 1 is also highly skewed, with most participants reporting low levels of symptoms severity (Fig. 3 and Fig. 2).

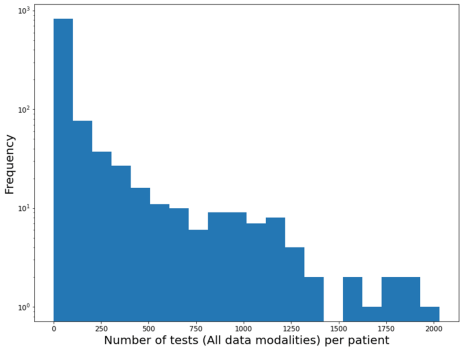

Fig. 1. The distribution of tests per patient, in logarithmic scale. Most participants contributed with a few tests.

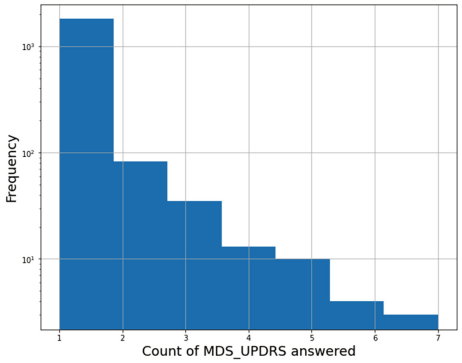

Fig. 2. Frequency of number of times the MDS_UPDRS questionnaire was answered by unique participants, in logarithmic scale. Most participants (N = 1951) answered this questionnaire only once during the study.

The skewness of these distributions is indicative of the fact that the majority of participants in the study were in relatively good health, were engaged in the study for a short time, and contributed to the study by completing a few tests and questionnaires. This was also observed in [13]. Training machine learning algorithms under these conditions is challenging and requires a well-designed pre-processing and filtering strategy.

3.3 Data Filtering and Pre-processing

As a first step, we collected the features available in the mPower dataset that corresponded to motor tasks and included additional information to link data to subjects through their 'healthCode' (unique subject identifier), 'createdOn' (timestamp of the data when it was recorded) and 'PD' (boolean variable indicating if the subject declared to have been diagnosed with PD). This yielded the following number of features for each activity type as seen in Table 2:

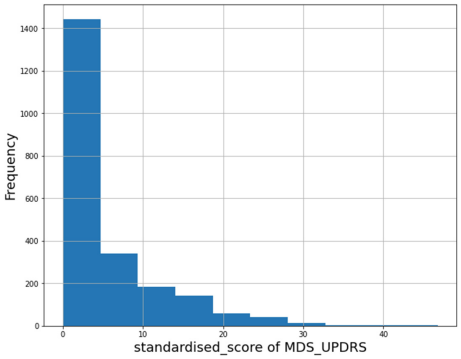

Fig. 3. Standardised score of MDS_UPDRS questions. The score is skewed towards lower values.

Table 2. The number of features for each activity type.

	Finger tapping	Walking	Rest	Voice
Nr. of features	45	116	22	16

After selecting motor features, we checked for missing values. There were missing values in some features across the different activity types, with the highest number of missing values in the 'PD' feature (self-declared PD diagnosis). As our analysis only focuses on actual PD patients, we decided to drop the rows where that value was missing and not try to impute them because we preferred to rely on accurate data.

In terms of motor symptoms level, we selected only those participants who answered the MDS-UPDRS questionnaire more than once. This is motivated by the risk in an uncontrolled data set like mPower - where anyone could download and use the app - that several participants wrongly declared themselves as diagnosed with PD. Without better control over the use and users, it is feasible that a number of users downloaded the app to try it, and, during that time, inserted fake data as they were testing how the app worked. Our hypothesis is, thus, that fake users would abandon the app quickly and that the data analysis would benefit from not including such users. As we observed that participants who responded the MDS-UPDRS questionnaire more than once had contributed with more tests, we used the number of answered MDS-UPDRS questionnaires as an indicator of participants' reliability.

After selecting participants we considered reliable, we used the answers to part 2 of the questionnaire to compute the normalized score and used it as our reference symptoms level. Activity data was then associated with symptoms level, by selecting the tests within +/− 14 days from the time each MDS-UPDRS questionnaire was answered. This was based on the hypothesis that motor symptoms, while known to be fluctuating every day, would not change if averaged within a 2-week period.

In order to account for the highly skewed distribution of performed activities per patient (standard deviation of 108.89), we limited the number of activities per patient to the 50th percentile computed on the whole population (92 for Finger tapping, 97 for Walking, 80 for Rest, 85 for Voice).

All features were normalized with the PowerTransformer from the scikit-learn library. This was used to make the data more 'Gaussian-like' to minimise the skewness of the distribution of each feature [14].

Finally, the most relevant features for each activity were selected. Since we are addressing this as a regression problem, we employed a backward elimination regression technique with a linear model [15]. After feature selection, we were left with 30 features for tapping data, 60 for walking data, 11 for rest data and 12 for voice data (including meta-information such as patient ID, timestamp and reference symptoms level).

Features Collapsing Strategy. When more than one activity of the same type (finger tapping, voice, walking, rest) was found within a time window of ± 14 days from the date of the reference symptoms level (derived from the answers to the MDS-UPDRS questionnaire), we computed the mean for each associated feature, grouping by participant, the symptoms level score and timestamps. This strategy has been used in previous research [13] to improve generalization and reduce the impact of identity confounding.

After averaging the features overlapping in the same time window, we merged all the features from all activities into one table together with timestamps and reference symptoms level. The resulting table contains, for each symptoms level, only one row with columns corresponding to the different features of the different activity types. Rows with missing columns, e.g. because of missing activity associated to a given symptoms level, were discarded.

Data Splitting Strategy. Similarly to [13], when splitting the data into training, test, and validation sets, we randomly shuffled the rows based on the participant identifier ('healthCode'). This was done to reduce the impact of identity confounding, which is strong in this dataset. The ratio between sets was 80/20 % between training and testing. When a validation set was required, the set was obtained from the training set by splitting the participants into 85/15 % for training and validation, respectively.

Regression Analysis. Given that our target attribute is a continuous attribute ranging between 0–52, we treated the problem of predicting the normalized symptoms level score as a regression problem. For that, we used 5 Machine Learning algorithms: Linear Regression, Lasso regression, Random Forest regressor, Support Vector Machine (SVM) and TabNet neural networks. Considering the number of data points we were left with after all the pre-processing steps, overfitting is a concern, thus, simpler models were applied such as Linear regression and Lasso regression. Furthermore, not having a huge dataset to feed, but having a high dimensional space after combining all the different data modalities, models such as SVM were supposed to perform well. TabNet was also tried

as a promising, though more computationally expensive, alternative for tabular learning [16].

A Monte Carlo cross-validation was used, where we re-shuffled patients for each of the training/test and validation sets randomly 5 times [17] and then averaged the 5 results for each evaluation metric. The shuffling was done based on subject ID so that the same subject could not end in both the training and the testing set. This was done to avoid the possibility of the model to learn more about specific subjects.

Evaluation Metrics. In order to evaluate our prediction analysis of the regression models, we used the following evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Adjusted R-squared score, Pearson's correlation and Spearman's correlation.

4 Results and Discussion

The effect of pre-processing and filtering data results in the data reduction shown in Table 3. Only 293 rows were eventually used in the regression analysis. These correspond to a total of 101 participants, 80 of which were randomly chosen for the training set, and 21 for the test set. When a validation set was required, 12 patients were removed from the training set.

Table 3. Number of participants and observations per activity after each step of the pre-processing and filtering pipeline.

	Finger tapping	Walking	Rest	Voice
After dropping missing values and non-PD				
Observations	42549	23391	23998	39051
Participants	1060	640	658	968
After selecting patients with >1 reported symptoms level and observations ± 14 days apart from symptoms level				
Observations	15444	12324	12819	14756
Participants	125	102	116	124
After limiting the number of tests per patient to 50th percentile				
Observations	8305	6832	6096	7512
Participants	125	102	116	124
After averaging features				
Observations	354	295	314	352
Participants	125	102	116	124
After collapsing features				
Observations	293	293	293	293
Participants	101	101	101	101

The performances of the algorithms employed in our analysis are shown in Table 4. All the algorithms achieve similar performances, with linear regression and lasso being slightly better and showing moderate to strong correlation between predicted and regressed values (a scatter plot for the linear regression algorithm is shown in Fig. 4). The algorithm obtaining the best performance is also the simplest, linear regression, whereas Tabnet, a deep-learning algorithm suitable for datasets with a considerably higher number of rows, obtains the worst metrics.

In terms of clinical applicability, the correlation between regressed symptoms level and the reference confirms the validity of the approach (i.e. what is measured is related to motor symptoms) [18]. The metrics refer to patients who were never introduced to the algorithm before, which suggests that the algorithms generalise well. However, none of the error statistics (mean absolute error, or root mean squared error) is below the clinically significant smallest change for part 2 of UPDRS, estimated between 3.05 and 2.51 [19], which indicates that the predictions are not accurate enough.

Table 4. Evaluation metrics of the regression algorithms. The results represent the mean result of 5 random different splits.

Evaluation Metric	Linear Regression	Lasso Regression	Random Forest	SVM	TabNet
Mean Absolute Error	5.4	5.5	5.6	5.9	5.8
Root Mean Squared Error	6.6	6.7	7.0	7.1	7.5
Adjusted R-squared score	2.5	2.5	2.6	2.7	2.7
Pearson's correlation	0.6	0.5	0.3	0.4	0.2
Spearman's correlation	0.5	0.5	0.4	0.4	0.3

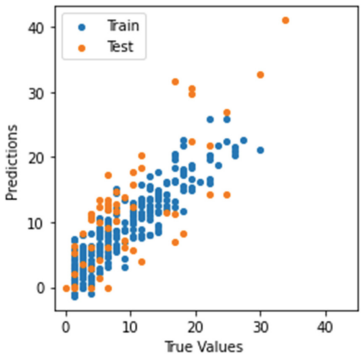

Fig. 4. Predicted symptoms level vs reference for the linear regression algorithm on the test and train sets. A positive correlation between the two quantities can be observed for the test and train set.

5 Conclusions

The mPower dataset contains an unprecedented quantity of data collected from mobile phones that can be used to detect and quantify Parkinson's disease symptoms. Given that the data was acquired in an uncontrolled manner, the dataset is skewed and likely to contain more noise. In this paper, we show how it is possible to process the dataset to focus on the parts of the dataset that is more reliable. Through such filtering, the number of participants was reduced from 1060 to 101 which we could confirm had contributed with high-quality data.

Using machine learning algorithms, we show that it is possible to correlate the data collected within the activities related to motor symptoms to the symptoms level, as measured from the answers to part 2 of the MDS-UPDRS questionnaire. The regressed level, however, still presents a high margin of error and should not be considered as a substitute for a conventional clinical assessment.

Future work could try to exploit the voluminous data available in mPower by exploring further optimization of the filtering stages with a goal to increase the number of remaining participants compared to our study and allowing more tests to be used in the training process in order to potentially improve accuracy. Additional studies could also aim at recruiting participants with a more uniform distribution across symptoms level compared with what the mPower data set currently shows, and ensuring that volunteers have been clinically diagnosed with PD.

Acknowledgment. This work was supported by the Mats Paulsson Foundation and the Internet of Things and People research center at Malmö University, funded by the Swedish Knowledge Foundation. These data were contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse [doi:10.7303/syn4993293].

References

1. Sveinbjornsdottir, S.: The clinical symptoms of Parkinson's disease. J. Neurochem. **139**, 318–324 (2016)
2. Dorsey, E.R., et al.: Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol. **17**(11), 939–953 (2018)
3. Goetz, C.G., et al.: Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov. Disord. Official J. Mov. Disord. Soc. **23**(15), 2129–2170 (2008)
4. Evers, L.J., Krijthe, J.H., Meinders, M.J., Bloem, B.R., Heskes, T.M.: Measuring Parkinson's disease over time: the real-world within-subject reliability of the MDS-UPDRS. Mov. Disord. **34**(10), 1480–1487 (2019)
5. Linares-Del Rey, M., Vela-Desojo, L., Cano-de La Cuerda, R.: Mobile phone applications in Parkinson's disease: a systematic review. Neurología (English Edition) **34**(1), 38–54 (2019)
6. Bot, B.M., et al.: The mpower study, parkinson disease mobile data collected using researchkit. Sci. Data **3**(1), 1–9 (2016)

7. Su, D., et al.: Simple smartphone-based assessment of gait characteristics in Parkinson disease: validation study. JMIR Mhealth Uhealth **9**(2), e25451 (2021)
8. Kuosmanen, E., et al.: Smartphone-based monitoring of Parkinson disease: quasi-experimental study to quantify hand tremor severity and medication effectiveness. JMIR Mhealth Uhealth **8**(11), e21543 (2020)
9. Chaibub Neto, E.L.I.A.S., et al.: Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iPhone sensor data. In: Biocomputing 2016: Proceedings of the Pacific Symposium. World Scientific, 2016, pp. 273–284 (2016)
10. Lauraitis, A., Maskeliūnas, R., Damaševičius, R., Krilavičius, T.: A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment. Sensors **20**(11), 3236 (2020)
11. Aghanavesi, S., Nyholm, D., Senek, M., Bergquist, F., Memedi, M.: A smartphone-based system to quantify dexterity in Parkinson's disease patients. Inform. Med. Unlocked **9**, 11–17 (2017)
12. Sieberts, S.K., et al.: Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's disease digital biomarker dream challenge. NPJ Digit. Med. **4**(1), 1–12 (2021)
13. L. Omberg, E., et al.: Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat. Biotechnol. **40**(4), 480–487 (2022)
14. Yeo, I.-K., Johnson, R.A.: A new family of power transformations to improve normality or symmetry. Biometrika **87**(4), 954–959 (2000)
15. Seabold, S., Perktold, J.: Statsmodels: econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
16. Arik, S.Ö., Pfister, T.: Tabnet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687 (2021)
17. Dubitzky, W., Granzow, M., Berrar, D.P.: Fundamentals of Data Mining in Genomics and Proteomics. Springer Science & Business Media, New York (2007). https://doi.org/10.1007/978-0-387-47509-7
18. Heale, R., Twycross, A.: Validity and reliability in quantitative studies. Evid. Based Nurs. **18**(3), 66–67 (2015)
19. Horváth, K., et al.: Minimal clinically important differences for the experiences of daily living parts of movement disorder society-sponsored unified Parkinson's disease rating scale. Mov. Disord. **32**(5), 789–793 (2017)