Building a Machine-learning Framework to Remotely Assess Parkinson's Disease Using Smartphones

Oliver Y. Chén, Florian Lipsmeier, Huy Phan, John Prince,
Kirsten I. Taylor, Christian Gossens, Michael Lindemann, and Maarten de Vos

***Abstract*—*Objective:* Parkinson's disease (PD) is a neurodegenerative disorder that affects multiple neurological systems. Traditional PD assessment is conducted by a physician during infrequent clinic visits. Using smartphones, remote patient monitoring has the potential to obtain objective behavioral data semi-continuously, track disease fluctuations, and avoid rater dependency. *Methods:* Smartphones collect sensor data during various active tests and passive monitoring, including balance (postural instability), dexterity (skill in performing tasks using hands), gait (the pattern of walking), tremor (involuntary muscle contraction and relaxation), and voice. Some of the features extracted from smartphone data are potentially associated with specific PD symptoms identified by physicians. To leverage large-scale cross-modality smartphone features, we propose a machine-learning framework for performing automated disease assessment. The framework consists of a two-step feature selection procedure and a generic model based on the elastic-net regularization. *Results:* Using this framework, we map the PD-specific architecture of behaviors using data obtained from both PD participants and healthy controls (HCs). Utilizing these atlases of features, the framework shows promises to (a) discriminate PD participants from HCs, and (b) estimate the disease severity of individuals with PD. *Significance:* Data analysis results from 437 behavioral features obtained from 72 subjects (37 PD and 35 HC) sampled from 17 separate days during a period of up to six months suggest that this framework is potentially useful for the analysis of remotely collected smartphone sensor data in individuals with PD.**

***Index Terms*—Parkinson's disease, remote disease assessment, feature-selection, machine-learning, predictive modeling, $P \gg N$ problem**

Manuscript received October 1, 2019; revised February 7, 2020; accepted April 7, 2020. Date of publication XXX XX, XXXX; date of current version XXX XX, XXXX. This work was funded by F. Hoffmann-La Roche Ltd and NIHR Oxford Biomedical Research Centre (BRC). *(Corresponding author: O. Y. Chén.)*

O. Y. Chén is with the Institute of Biomedical Engineering (IBME), University of Oxford, Oxford OX3 7DQ, U.K. (email: yibing.chen@seh.ox.ac.uk).

F. Lipsmeier, C. Gossens, and M. Lindemann are with Roche Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland.

H. Phan is with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K.

J. Prince was with IBME, University of Oxford, Oxford, U.K.

K. I. Talor is with Roche Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland; Faculty of Psychology, University of Basel, Basel, Switzerland.

M. de Vos is with IBME, University of Oxford, Oxford, U.K.; Department of Electrical Engineering and Department of Development and Regeneration, KU Leuven, Leuven, Belgium.

M. de Vos and M. Lindemann contributed equally.

I. Introduction

Parkinson's disease (PD) affects seven million people worldwide; the prevalence increases from 1% of the population for those over 60 years of age to 4% over 80 [1]. A reliable, objective, fast, and remote method to quantify the presence and severity of PD symptoms would benefit a large number of people who are affected by, or are at risk to develop, PD.

Previous studies have measured common PD symptoms with object- and technology-based tests, such as sustained phonation (*i.e.* voice) [13], [14], rest tremor [15]–[17], postural tremor [18], [19], dexterity [11], [20], balance [21]–[23], and gait [22], [24]. Advancement in digital technologies makes data collection using smartphones increasingly convenient and accurate. Smartphones are small, portable, and widely-used. The data captured from various smartphone sensors can be remotely transferred via wireless networks, facilitating out-clinic data collection and assessment. Because of these attractive properties, researchers have begun to explore the possibilities of studying PD using smartphone data, and have brought in new avenues to remote PD assessment [2]–[12].

In spite of these promises, remote PD assessment using smartphones is still in its infancy. **Table** 1 gives an overview of recent PD studies using machine-learning approaches on smartphone features. Although existing methods and analyses have used different datasets with various sample sizes, the overview shows that, in general, studies have considered few and inconsistent feature modalities, and reported performance accuracy via varying statistical approaches. Additionally, most models were developed with a relatively limited scope that was either restricted to disease classification or disease severity estimation. Moreover, some studies only considered PD samples. Here, in light of existing efforts, we propose a unified machine-learning framework that (1) extracts disease- or symptom-specific features from a rich variety of sensor data, (2) takes into account the differences between PD participants and HCs, (3) builds the selected features into a relevant feature map, (4) differentiates PD cases from HCs, and (5) estimates disease severity.

The framework first employs a two-step feature selection procedure and identifies features that are potentially associated with the disease (in terms of diagnostic group or severity). The selected features then enter the elastic-net regularized regression model to construct a feature map consisting of parameter estimates. Subsequently, the model links the feature map with

Recent studies	Sample size (PD/HC)	Number of repetitions	Out clinic	Modalities considered							Accuracy	Ensemble† improvement
				Voice	Gait	Balance	Dexterity	Rest tremor	Postural tremor	Others		
Current study	37/35	$4,883^*$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	0.973/0.971 (Sens/Spec) 0.987 (Accuracy) 0.993 (AUC)	Yes
Prince *et al.* (2018) [2]	949/866	NA	Yes	No	No	No	Yes	No	No	No	0.65 (Accuracy)	No
Zhan *et al.* (2018) [3]	129/0	6,148	Yes	Yes	Yes	Yes	Yes	No	No	Reaction Time	0.81 (Pearson correlation)	Yes
Prince *et al.* (2018) [4]	312/236	48,892	Yes	No	No	No	Yes	No	No	Memory	NA	No
Bot *et al* (2016). [5]	1087/5581	78,887	Yes	Yes	Yes	No	Yes	No	No	Memory	NA	No
Zhan *et al.* (2016) [6]	121/105	1,600	Yes	Yes	Yes	Yes	Yes	No	No	Reaction Time	0.693/0.727 (Sens/Spec)	Yes
Neto *et al.* (2017) [7]	23/23	NA	Yes	Yes	Yes	No	Yes	No	No	No	0.5-0.6 (AUC)	No
Arora *et al.* (2015) [8]	10/10	18	Yes	Yes	Yes	Yes	Yes	No	No	Reaction Time	0.962/0.969 (Sens/Spec)	No
Lee *et al.* (2016) [9]	57/87	432	No	No	No	No	Yes	No	No	No	0.92 (AUC)	No
Arroyo-Gallego *et al.* (2017) [10]	21/23	51	No	No	No	No	Yes	No	No	No	0.810/0.810 (Sens/Spec)	No
Kassavetis *et al.* (2016) [11]	14/0	14	No	No	No	No	Yes	No	No	No	NA	No
Printy *et al.* (2014) [12]	18/0	54	No	No	No	No	Yes	No	No	No	NA	No

Table 1: An overview of PD studies using smartphone features. We selected twelve recent and representative studies that used smartphone data to study PD. We listed key characteristics, including the sample size, the total number of repetitions, whether the study was conducted outside of clinics, the type of tests used, the estimation accuracy (if any), and ensemble improvement. Whether the study was conducted outside of clinics is important because collecting measurements frequently in clinics is inconvenient for large-scale examination in practice. *The repetition means the total number of data points across all features and subjects. In other words, if the j^{th} $(1 \leq j \leq P)$ feature of subject i $(1 \leq i \leq N)$ was measured over T_{ij} days, the repetition is $\sum_{i=1}^{N} \sum_{j=1}^{P} T_{ij}$. †If *ensemble improvement* equals to yes, it means that using cross-modality features (*i.e.* features obtained from different behaviors) improves the estimation accuracy. Shaded orange *vs.* grey color indicates if a study covers a specific component.

features from the training subjects to estimate their diagnostic group status and severity. To evaluate the reproducibility of the framework, the model tests the feature map on features from novel (testing) subjects to perform out-of-sample PD assessment. The proposed framework is illustrated in **Figure** 1.

We arrange the rest of the article as follows. In **Section** II, we introduce the smartphone data used in this study. In **Section** III-A, we define notations and describe data organization. In **Section** III-B, we provide the main methodological framework and its building blocks. **Section** III-C highlights the framework's applications in PD/HC classification and PD severity estimation. In **Section** IV, we present experimental and data analysis results. We discuss future work in **Section** V and conclude the article in **Section** VI.

II. The Home-based PD Data Collected by Smartphones

We use data collected from two independent smartphone-based remote monitoring studies [25]. The first study was a six-month-long phase 1b clinical drug trial of PRX002/RG7935 (now known as prasinezumab) conducted by Prothena and Roche, which consisted of 44 PD participants (NCT02157714). The second study was a six-week-long observational study of 35 age- and sex-matched healthy controls (HCs). The respective local ethics committees approved both studies. Written informed consent was obtained from all participants (patient study: IRB00010809, H-35018, WOR1-14-143; control study: EKNZ-BASEC-2016-00596). All controls scored $\geq$ 26 points on the Montreal Cognitive Assessment (MoCA) [26] and were free of cardiovascular, neurological or psychiatric condition, and had no first-degree relative with PD. The study also included the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). MDS-UPDRS scores measure the progression of an individual's Parkinson's disease, and serve as the gold standard for validation [27]. Throughout, we used the MDS-UPDRS total scores. The total score equals to the summation of sub-scores obtained from 42 items covering four subscales: Part I: mentation, behavior, and mood; Part II: activities of daily living; Part III: motor examination; Part IV: complications of therapy. The total score ranges from 0 to 199 points, where a patient with a higher score would be considered to have more severe PD [28]. In this study, the scores were administered by trained raters (Parts I and II) and physicians (Part III) to subjects during screening (study days -42 to -1) and days 8 and 64. Trained raters tested controls at baseline and day 42.

Both the PD and HC studies followed identical procedures. During the initial in-clinic visit, all subjects received a smartphone (Galaxy S3 mini; Samsung, Seoul, South Korea) with the Roche PD Mobile Application v1 (Roche, Basel,

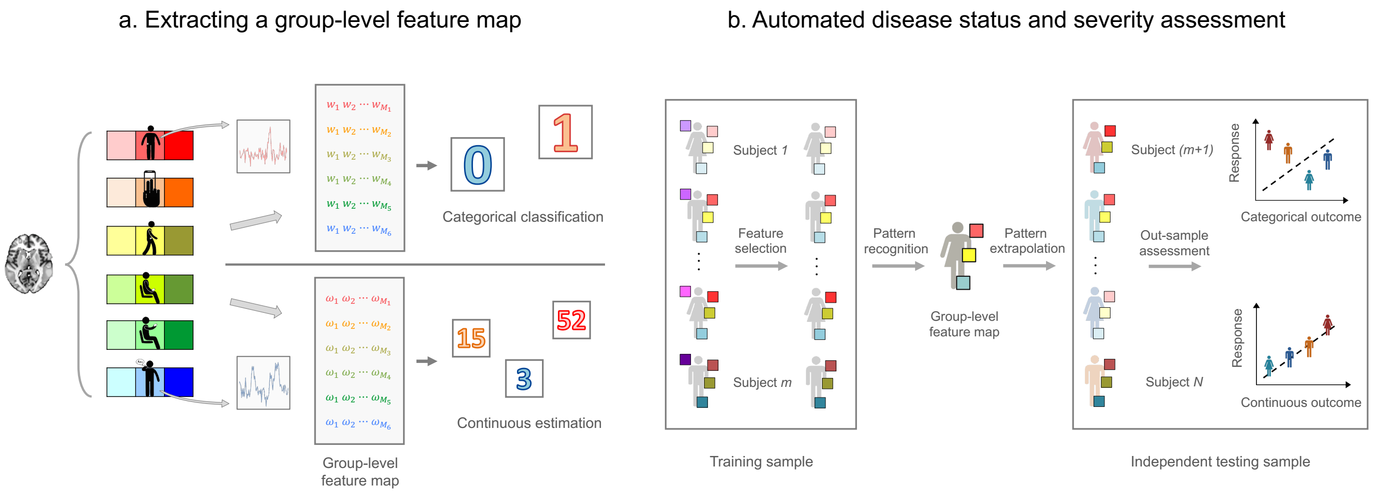
a. Extracting a group-level feature map
b. Automated disease status and severity assessment
$w_1\ w_2 \cdots w_{M_1}$
$w_1\ w_2 \cdots w_{M_2}$
$w_1\ w_2 \cdots w_{M_3}$
$w_1\ w_2 \cdots w_{M_4}$
$w_1\ w_2 \cdots w_{M_5}$
$w_1\ w_2 \cdots w_{M_6}$
0
1
Categorical classification
$\omega_1\ \omega_2 \cdots \omega_{M_1}$
$\omega_1\ \omega_2 \cdots \omega_{M_2}$
$\omega_1\ \omega_2 \cdots \omega_{M_3}$
$\omega_1\ \omega_2 \cdots \omega_{M_4}$
$\omega_1\ \omega_2 \cdots \omega_{M_5}$
$\omega_1\ \omega_2 \cdots \omega_{M_6}$
15
52
3
Continuous estimation
Group-level feature map
Subject 1
Feature selection
Subject m
Training sample
Pattern recognition
Group-level feature map
Pattern extrapolation
Subject (m+1)
Out-sample assessment
Subject N
Response
Categorical outcome
Response
Continuous outcome
Independent testing sample

Figure 1: (a) Extracting a group-level feature map. Each color specifies a feature modality. Boxes with the same color but with different hues indicate multiple behavioral features from the same feature modality. The red, orange, yellow, chartreuse green, and blue boxes refer to balance, dexterity, gait, rest tremor, postural tremor, and voice, respectively. The distinctive performance patterns on these tasks correspond to the functioning of specific functional-neuroanatomic circuits, which cannot be directly assessed (indicated by a gray bracket). The model couples behavioral features with a trained group-level feature map, yielding estimated outcomes. The colored Latin and Greek letters with subscripts (*e.g.* w_1 and ω_1) represent feature weights across features. (b) Automated disease group and severity assessment. During the model building step, features that are relevant to the targeted outcome are selected. Subsequently, a feature map consisting of weights across selected features is developed using data from individuals in the training sample. The weights indicate how to integrate features to yield an estimation for the targeted, discrete or continuous, outcome. During the prospective testing step, the efficacy of the feature map is verified by applying the map to features from previously unseen individuals without further model fitting, which yields estimations for each subject. The model produces one estimated outcome (binary disease group or continuous disease severity) per subject. The consistency of the features and the reproducibility of the model can then be evaluated by comparing the observed and estimated outcomes in the testing sample. For binary classification, we report statistics such as accuracy, *kappa*, sensitivity, and specificity. For continuous disease score estimation, we report RMSE and correlation between estimated and observed disease scores as measured by the MDS-UPDRS.

Switzerland) preinstalled. They also received a belt containing a pouch that carried the phone. Smartphones were "locked-down" (*i.e.* configured so patients could only run the Roche PD Mobile Application v1 and WiFi connection software). Site staff provided the subjects training on the active tests. Subsequently, subjects were instructed to complete the active tests at home once daily (in the morning), to carry the phone with them throughout the day, and to recharge the phone overnight.

A full description of the study and data processing can be found in [25].

III. Methods

A. Notations and Data Organization

We begin by defining the notations used throughout this article. To ensure that the estimation power is not influenced by the amount of data that was available to each individual, unless otherwise specified, we truncate the raw data such that every subject has data from the same number of days (17 in our study). A thorough treatment of missing data, such as imputation, is available elsewhere [29], [30].

Let N denote the number of subjects in the study, where $N = 72$. The i^{th} subject, for $1 \le i \le N$, has T days, where $T = 17$. During each day, features from K modalities are measured, where $K = 6$ in the study. Each modality contains further features. Specifically, the k^{th} $(1 \le k \le K)$ modality contains M_k features, where M_k ranges from 37 to 178. The m^{th} feature of the k^{th} modality, is measured at time points $1, 2, \ldots, T$, for the i^{th} subject during the j^{th} day. Thus, each feature takes the form $x_{ikm}(t_j)$, for $1 \le i \le N$, $1 \le k \le K$, $1 \le m \le M_k$, and $1 \le t_j \le T$. Let $\sum_{k=1}^{K} M_k = P$, where $P = 437$ in the study. That is, there are a total of P features. Thus, the feature data $\mathbf{X}$ is a data cube of size $N \times P \times T$.

Similarly, we denote the outcome as $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, where y_i, for $1 \le i \le N$, is a categorical label in case of binary classification (*i.e.* PD *vs.* HC) and a continuous value in case of PD severity estimation (*i.e.* MDS-UPDRS total score).

To discover features useful for estimating an outcome, we first summarize each feature by their first moment (arithmetic mean). Formally, the first moment of the m^{th} feature from the k^{th} modality of the i^{th} subject is defined as $\xi_{ikm} = \frac{\sum_{t_j=1}^{\tilde{T}_i} x_{ikm}(t_j)}{\tilde{T}_i}$, where $\tilde{T}_i$ indicates the number of days during which features are averaged for each individual.

Throughout the article, we use the first moment approach to summarize features for model building, because the mean conveniently provides the fundamental information of the features. In **Section** V, we will discuss advantages and limitations of using the mean to summarize the features.

B. Machine-learning Framework

Our framework consists of two parts: (1) feature selection; (2) model building and automated disease assessment.

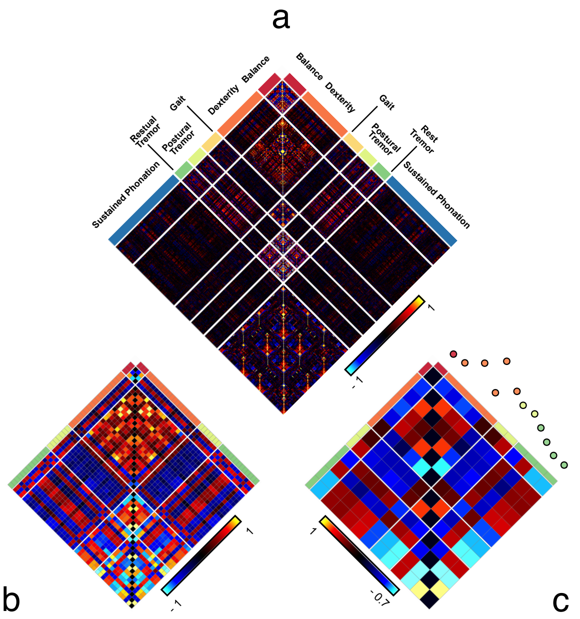

Figure 2: An example of a two-step feature selection procedure. (a) A 437 × 437 matrix consisting of pair-wise Pearson correlations between 437 features obtained from six behavior modalities. The features are more correlated within their respective modality than they are with features out of that modality. (b) A 41 × 41 matrix consisting of pair-wise Pearson correlations between 41 features, selected from step-one feature selection. (c) A 12 × 12 matrix consisting of pair-wise Pearson correlations between 12 features, selected from step-two feature selection. The height of each colored circle above the heatmap indicates the weighted contribution (in the sense the higher the more important in terms of disease assessment) of each corresponding feature. Here, *weighted* means the features are scaled (mean 0, standard deviation 1) so that the magnitude of the weights is not biased by features with large means or variances.

1) Feature Selection: For a $P \gg N$ *problem* (also known as the "short, fat data problem", where the number of features P is much larger than the number of samples N), there are commonly two difficulties. First, it suffers from the "curse of dimensionality", where the curse is twofold: (i) the number of samples (N) needed to yield a reliable statistical result grows exponentially as P grows, but a large number of sample is usually difficult to obtain; (ii) when P is large, all subjects appear to be dissimilar. This makes extracting common features (*e.g.* features shared by PD participants or HCs) difficult. Second, there is an insufficient number of degrees of freedom to estimate the full model. To alleviate these challenges, we demonstrate that only a small number of the P features are needed to build the model (see **Section** IV-D for details). This assumption is supported, in part, by the strong correlation observed in the PD feature data, which we illustrate below in detail.

Some of the PD sensor feature data are strongly correlated; the intra-modality features are more correlated than inter-modality features (see **Figure** 2). When several highly correlated features are associated with an outcome, choosing one of them is analytically sufficient, and will give the most parsimonious model. The discarded (relevant) features, however, may uncover an underlying property that is meaningful for interpreting the biological system. For example, consider two modalities, say voice and tremor, each with 100 features. Suppose there are 10 highly correlated voice features and 50 highly correlated tremor features that are associated with the disease outcome, and, for simplicity, suppose voice features are not highly correlated with tremor features. A model built on 1 (out of the 10 selected) voice feature and 1 (out of the 50 selected) tremor feature, therefore, is as sufficient as a model built using all selected features. The discarded features, however, may offer a better (and easier) biological explanation for the outcome. In this regard, it is important to consider a model that can account for both parsimony (*i.e.* removing redundant features) and biological interpretation (*i.e.* allowing a few correlated features).

To that end, we introduce a two-step feature selection procedure tailored for large-scale data. During the first step, we eliminate features that are not significantly related to the outcome (in the training data) using a mass-univariate approach.

For a continuous outcome (in our study the MDS-UPDRS total score), the first step involves a feature-wise correlation test to examine whether or not each feature is significantly correlated with the MDS-UPDRS total score, using a correlation test. Since the overall model incorporated an identity link function for continuous outcome assessment (coupled with the elastic-net, a regularized linear model), we used Pearson correlation test to identify features that were linearly associated with the continuous outcome. Although the selected features are significantly correlated with the outcome to various degrees, they are not necessarily significantly correlated with each other (see **Figure** 2). The heterogeneous groups of features, therefore, may each address a proportion of variability of the outcome.

For a binary, or categorical, outcome (in our study the binary disease status), since a correlation test is inappropriate, the first step involves a feature-wise *t*-test to examine whether or not each feature varies significantly across groups.

During the second step, the selected features are further pruned via regularization. Common regularization approaches include the Lasso (least absolute shrinkage and selection operator) regularization [31] and the Ridge regularization (or the Tikhonov regularization) [32]. The Lasso picks one feature among all correlated ones, on which a single non-zero weight is imposed, whereas the Ridge imposes weights on all correlated features and, then, averages their coefficients in order to reduce the effect of multiple correlated features to the full model.

2) Model Building and Automated Disease Assessment: Chief to automated disease assessment when the number of features (denoted as P) is much larger than the number of samples (denoted as N) is a modeling technique called regularization. A regularized model, such as Lasso and Ridge, shrinks the estimated parameters of irrelevant features (and therefore suggests either removing, or punishing the weights of, these features in the model output). The elastic-net regularization combines the Lasso and the Ridge regularizations, and offers a compromise between them [33]. It chooses a small number of features (like the Lasso), some of which are correlated (like the Ridge), which may provide useful biological interpretation of PD data. Because of its balance between interpretability and parsimony, we use the elastic-net during the second-step feature selection.

Consider a feature $\boldsymbol{\xi}$. Denote $\rho_{(\boldsymbol{\xi},\mathbf{y},N-2)}$ as the result from a statistical test during the step-one feature selection between the feature $\boldsymbol{\xi}$ and the outcome $\mathbf{y}$. The value of ρ can be a t-statistic from a t-test for a binary outcome or a correlation for a continuous outcome; equivalently, it could be the corresponding p-values. Let ϵ be a pre-specified threshold for ρ. Although there is a one-to-one mapping between a statistic and its p-value, sometimes it may be convenient to evaluate the p-value, whereas other times it may be convenient to evaluate the t-statistic or the correlation. For example, in this study we threshold the t-statistic of binary t-tests at 5 during disease classification, and threshold the p-value of correlation tests at 0.01 during disease severity estimation.

Formally, we define our model as

$$\begin{aligned}\mathbb{E}(y_i\,|\,\boldsymbol{\xi}_i,\boldsymbol{\delta}_i) = g^{-1}(\mu + \mathbf{f}_i^{\intercal}\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^{\intercal}\boldsymbol{\gamma}) \\ +\lambda_2|\mathbf{P}|_2^2 + \lambda_1|\mathbf{P}|_1\end{aligned} \tag{1}$$

where $g(\cdot)$ is a link function, μ is the intercept, $\mathbf{f}_i^{\intercal} = [\,\mathbf{f}_{i1}^{\intercal}, \mathbf{f}_{i2}^{\intercal}, \ldots, \mathbf{f}_{iK}^{\intercal}\,]$, $\mathbf{f}_{ik}^{\intercal} = (\xi_{ik1}, \xi_{ik2}, \ldots, \xi_{ikM_k})$. Recall that there are K total modalities, with the k^{th} modality containing M_k features. Here, $\mathbf{S} = \mathit{blockdiag}\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_K\}$, and $\mathbf{I}_k = \text{diag}\{i_{k1}, i_{k2}, \ldots, i_{kM_k}\}$, wherein $i_{km} = 1$ if $\rho_{(\boldsymbol{\xi}_{km},\mathbf{y},N-2)} < \epsilon$, and 0 otherwise, where $\boldsymbol{\xi}_{km} = (\xi_{1km}, \xi_{2km}, \ldots, \xi_{Nkm})$ (a particular feature across all subjects); $\boldsymbol{\delta}_i$ is a vector containing all covariates for the i^{th} subject (in this study the covariates are age and gender), and $\boldsymbol{\gamma}$ is its coefficient; $\mathbf{P} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\gamma}]^{\intercal}$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kM_k})$, for $k = 1, 2, \ldots, K$. Finally, λ_1 and λ_2 are penalty parameters.

Algorithm 1: A generalized two-step feature selection and predictive framework for automated disease assessment

```
Step 0: Reshape $\mathbf{X}$ to be of size $N \times P \times T$, where $P = \sum_{k=1}^{K} M_k$.
Step 1: For every feature $m$ of modality $k$ of subject $i$, compute the temporal mean $\xi_{ikm}$. We stack all subject's mean feature as an $N \times P$ matrix $\mathbf{F}^{\intercal}$.
Step 2: Conduct the step-one feature selection to obtain estimate $\hat{\mathbf{S}}$ of $\mathbf{S}$ in Eq. (1). The selected features are then $\mathbf{F}^{\intercal}\hat{\mathbf{S}}$.
Step 3: Conduct the step-two feature selection via (the elastic-net) regularization. The remaining features are those whose estimated parameters in Eq. (2) are non-zero.
Step 4: Run out-of-sample disease assessment using estimates from Eq. (1).
```

Through standard linear algebraic manipulation [33], the solution for Eq. (1) is

$$\hat{\boldsymbol{\beta}} = \sqrt{1+\lambda_2}\left\{ \operatorname*{arg\,min}_{\boldsymbol{\beta}^*} |\mathbf{y}^* - \mathbf{Z}^*\boldsymbol{\beta}^*|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}|\boldsymbol{\beta}^*|_1 \right\} \tag{2}$$

where $\mathbf{y}^*_{n+p} = \{g\left(\mathbb{E}(y_1\,|\,\boldsymbol{\xi}_1,\boldsymbol{\delta}_1)\right), g\left(\mathbb{E}(y_2\,|\,\boldsymbol{\xi}_2,\boldsymbol{\delta}_2)\right), \ldots, g\left(\mathbb{E}(y_p\,|\,\boldsymbol{\xi}_p,\boldsymbol{\delta}_p)\right), \mathbf{0}_p\}^{\intercal}$ and $\mathbf{Z}^*_{(n+p)\times p} = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathbf{F}^{\intercal}\mathbf{S}\\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix}$.

The choice of λ_1 and λ_2 are determined in two steps: for each fixed λ_2, we find the optimal λ_1; subsequently we find the optimal λ_2 along the selected λ_1 [33]. When $\lambda_1 = 0$, Eq. (2) reduces to the Lasso solution; when $\lambda_2 = 0$, Eq. (2) reduces to the Ridge (or the Tikhonov) solution. Any other (elastic) choices of λ_1 and λ_2 form a compromise between the Ridge and the Lasso regularization. The compromise can be illustrated by rewriting the penalty terms in Eq. (1) as

$$(1-\alpha)|\mathbf{P}|_2^2 + \alpha|\mathbf{P}|_1 \tag{3}$$

where α is called a mixing parameter [33], which controls how much "Lasso-ness" and "Ridge-ness" the regularization chooses. Specifically, when $\alpha = 0$, the regularization is strictly Ridge, and when $\alpha = 1$, the regularization is strictly Lasso.

C. PD Assessment

In the following, we apply the framework outlined in Eq. (1) in two specific scenarios: (i) PD/HC classification, and (ii) PD severity estimation.

(i) PD/HC Classification. When the outcomes are binary (*e.g.* diseased *vs.* healthy), the link function in Eq. (1) is $g(x) = \ln(\dfrac{x}{1-x})$ (*i.e.* the inverse of logistic function). Formally,

$$\begin{aligned}P(y_i = 1\,|\,\boldsymbol{\xi}_i,\boldsymbol{\delta}_i) = \frac{\exp(\mu + \mathbf{f}_i^{\intercal}\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^{\intercal}\boldsymbol{\gamma})}{1+\exp(\mu + \mathbf{f}_i^{\intercal}\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^{\intercal}\boldsymbol{\gamma})} \\ +\lambda_2|\mathbf{P}|_2^2 + \lambda_1|\mathbf{P}|_1\end{aligned} \tag{4}$$

where i refers to the i^{th} subject. The estimated conditional disease propensity, or $P(y_i = 1\,|\,\boldsymbol{\xi}_i,\boldsymbol{\delta}_i)$, is further thresholded to be 1 if it is greater than 0.5, or 0 otherwise. The results are shown in **Section** IV-B

(ii) Estimation of PD severity. When the outcomes are continuous (*e.g.* the MDS-UPDRS total scores), the link function in Eq. (1) is $g(x) = x$ (*i.e.* an identity mapping). Formally,

$$\mathbb{E}(y_i\,|\,\boldsymbol{\xi}_i,\boldsymbol{\delta}_i) = \mu + \mathbf{f}_i^{\intercal}\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^{\intercal}\boldsymbol{\gamma} + \lambda_2|\mathbf{P}|_2^2 + \lambda_1|\mathbf{P}|_1 \tag{5}$$

where i refers to the i^{th} subject.

The results are shown in **Section** IV-C.

IV. Experiments and Results

A. Cross-Validation Setup and Model's Parameters

To evaluate the performance of the framework, we split the data from N subjects described in Section II into four folds and conducted four-fold cross-validation. We used four statistics (accuracy, *kappa*, specificity, and sensitivity) to evaluate binary disease classification performance (*i.e.* PD *vs.* HC); we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2020.2988942, IEEE Transactions on Biomedical Engineering

(a) Results of PD/HC classification

	Multivariate Logistic Regression		XGBoost	Elastic-net			
				Raw biomarkers		Mean biomarkers	
	One-step	Two-step		One-step	Two-step	One-step	Two-step
Accuracy	0.493	0.813	0.944	0.947	0.947	0.960	0.973
Kappa	0.0035	0.625	0.889	0.894	0.894	0.920	0.947
Specificity	0.629	0.800	0.943	1.000	1.000	1.000	1.000
Sensitivity	0.375	0.825	0.946	0.900	0.900	0.925	0.950
Misclassified Subjects	38	14	4	4	4	3	2
Computing Time	2.46 mins	2.76 mins	<1min	40.26 mins	42.23 mins	3.09 mins	3.31 mins

(b) Results of PD severity assessment

		Multivariate Logistic Regression		XGBoost	Elastic-net			
					Raw biomarkers		Mean biomarkers	
		One-step	Two-step		One-step	Two-step	One-step	Two-step
Both PD and HC data	RMSE	503.21	553.53	23.21	602.31	48.59	23.88	16.58
	Pearson correlation	0.12	0.28	0.67***	-0.08	0.41***	0.57***	0.72***
	Computing time	<1 min	<1 min	<1 min	71 min	14 min	<1 min	<1 min
Only PD data	RMSE	1927.66	2827.81	30.81	567.55	71.16	27.66	17.19
	Pearson correlation	-0.14	0.30	0.11	-0.26	0.42**	0.16	0.54 ***
	Computing time	<1 min	<1 min	<1 min	37 min	4 min	<1 min	<1 min

Table 2: Results of PD/HC classification and disease severity assessment. We compared the performance of our framework with it using the baseline approaches. All results were cross-validated; RMSE refers to the root mean square error and ** and *** indicate that the Pearson correlations were significant at $p < 0.01$ and $p < 0.001$, respectively. The computing time was calculated using a Macintosh computer with 2.4 GHz Intel Core i5 processor.

used RMSE and correlation between observed and estimated outcomes to evaluate continuous PD severity assessment performance. The observed outcomes are individual mean (over time) MDS-UPDRS total scores and estimated outcomes are estimated mean MDS-UPDRS total scores.

We summarize the experimental set-up in Algorithm 1. The analyses were performed using the *R* software via customized codes. The second step feature selection was conducted using the elastic-net regularization provided by *R* package *glmnet* [34]. Two parameters were tuned for the two-step feature selection and predictive framework: ϵ, a threshold used during the first step of feature selection, and α ($0 \leq \alpha \leq 1$), a mixing parameter controlling how much Ridge-ness or Lasso-ness the elastic-net was. For disease classification, ϵ was used to threshold *t*-statistics and was set at 5; namely, a feature would be selected if its *t*-statistic was above 5. For disease severity estimation, ϵ was used to threshold *p*-values and was set at 0.01; namely, a feature would be selected if its *p*-value was below 0.01. We also provided the computing time needed to evaluate the model efficient on the same computer (a standard Macintosh computer with 2.4 GHz Intel Core i5 processor).

To demonstrate the efficacy of the proposed framework, we compared it to multivariate logistic regression (MLR) and XGBoost models in the same cross-validation strategy. To show the advantage of using mean features, we applied the proposed framework to the raw features (where repeated measurements of one feature are considered as multiple samples). We recorded the accuracy statistics from each of the alternative approaches in the following section, with a discussion.

B. Binary PD/HC Classification Results

In **Table** 2, as an initial step to understand the machine-learning framework we introduced in this article, we presented the model's performance on binary PD/HC classification using Eq. (4). There are three points to note. First, across multiple models, the two-step feature selection procedure yielded a higher estimation accuracy than a one-step feature selection procedure. Even with regularization, the two-step feature selection procedure still marginally improved accuracy and sensitivity. Second, using mean features significantly reduced computing time from 40 minutes (using raw features) to 3 minutes (using mean features), meanwhile improving estimation accuracy mildly. Third, our framework outperformed the baseline MLR and XGBoost models in identifying PD participants and HCs (see **Table** 2 (a)).

The disease assessment accuracy and the number of selected features depend on the mixing parameter α in Eq. (3). Nevertheless, across α values, a majority of selected features belong to dexterity and rest tremor modalities. Specifically, when setting $\alpha = 0$ (*i.e.* the Ridge), 38 out of the 53 final features are from dexterity and rest tremor modalities; when setting $\alpha = 1$ (*i.e.* the Lasso), 18 out of the 25 final features are from them (see **Figure** 3). The contribution each feature modality makes to the disease assessment is highlighted in **Figure** 2 (c), where dexterity shows the highest importance followed by rest tremor. Taken together, our results suggest the importance of dexterity and rest tremor features in PD assessment.

Although we showed that it is possible to identify PD participants from HCs using 17 non-contiguous days' of data with high accuracy, it remained unclear how many days of data are required to yield a stable estimate of the disease status. To check for minimal data requirement, we applied Eq. (4) to data obtained from an increasing number of days, and demonstrated that PD can be reliably identified using 10 non-contiguous days' behavioral data (see **Figure** 4).

In summary, our results suggest that (i) the two-step feature selection procedure generally outperforms more traditional approaches in classification accuracy; (ii) the mean approach

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2020.2988942, IEEE Transactions on Biomedical Engineering

is computationally more efficient than using raw features; (iii) behavioral data obtained in 10 non-contiguous days can reliably distinguish PD participants from HCs.

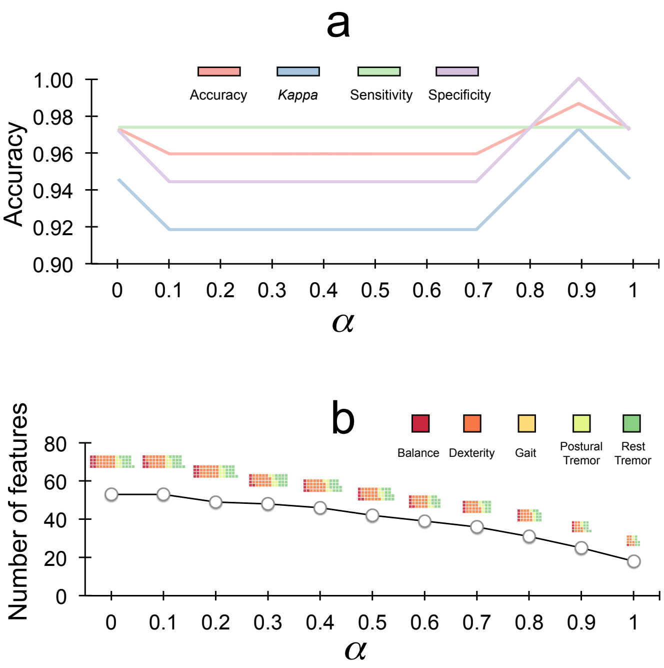

Figure 3: Estimation accuracy and number of features selected when the elastic-net mixing parameter α takes values from 0 to 1. (a) We examine four statistics (accuracy, *kappa*, sensitivity, and specificity) for evaluating model estimation accuracy. (b) When α increases, the number of selected features reduces. We use color code to uncover the distribution of features across each modality. Of note, when $\alpha = 0$, the elastic-net regularization reduces to the Ridge regularization; when $\alpha = 1$, the elastic-net regularization reduces to the Lasso regularization.

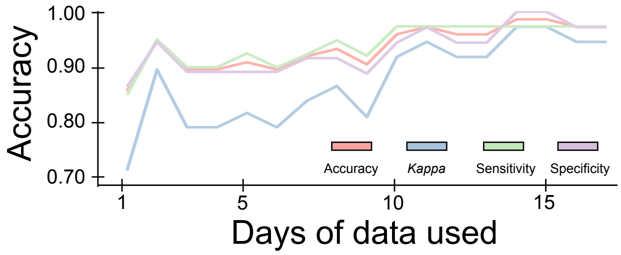

Figure 4: Determining the minimal amount of data needed to build a stable model. Each colored curve represents a function of how many days' of data are averaged. The results suggest that accuracy improves and stabilizes once more than 10 non-contiguous days' of data are used.

C. PD Severity Model Results

We carried out the assessment of continuous PD severity (*i.e.* the MDS-UPDRS total scores) using Eq. (5) in two experiments. During the first experiment, we conducted disease assessment using data from both PD participants and HCs. Note that (a) not all HCs' MDS-UPDRS total scores are 0; and (b) all MDS-UPDRS total scores are non-negative, and a large MDS-UPDRS total score indicates the disease is more severe. Our analysis showed a Pearson correlation of 0.72 $(p < 0.0001)$ and a RMSE of 16.58 between the estimated and the observed MDS-UPDRS total scores (see **Figure** 5 (a) and **Table** 2). Although the estimated MDS-UPDRS total scores in the novel samples are significantly correlated with the observed scores with a small RMSE, it remained possible that the high estimation accuracy was merely driven by obvious differences in features between the PD participants and HCs. To check for this possibility, we sought to perform an additional experiment. During the second experiment, we only used features from PD participants $(N = 37)$. The analysis yielded a significant (although smaller) Pearson correlation of 0.54 $(p < 0.001)$ and a similar RMSE of 17.19 (see **Figure** 5 (b) and **Table** 2).

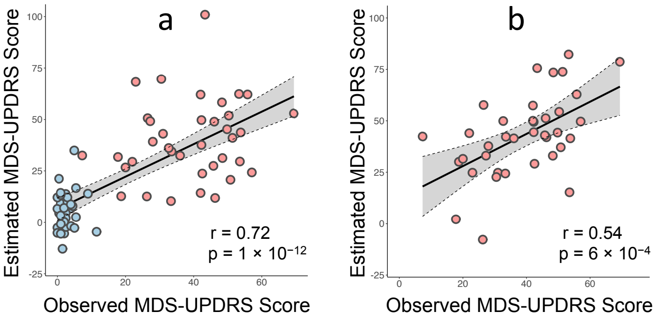

Figure 5: Assessing subjects' continuous MDS-UPDRS total score using individual features. (a) Estimation result of MDS-UPDRS total scores using data from both PD participants and HCs $(N = 72)$. (b) Estimation result of MDS-UPDRS total scores using data from PD participants only $(N = 37)$. Each dot represents one subject; gray area represents 95% confidence interval for best-fit line.

For PD severity assessment, a two-step feature selection procedure (namely, a step-one feature selection via a pairwise Pearson correlation test, followed by a step-two feature selection via the elastic-net regularization on selected features) yielded a higher estimation accuracy than a one-step feature selection procedure (namely, only using the elastic-net regularization) across multiple models. Additionally, using mean features significantly reduced computing time while improving estimation accuracy mildly. Finally, our framework outperformed the baseline MLR and XGBoost models in assessing continuous MDS-UPDRS total scores (see **Table** 2 (b)).

Taken together, for disease classification, the proposed framework selected features with heterogeneous profiles between PD participants and HCs. Further, for disease severity estimation, the framework reliably discovered features related strongly to continuous disease severity measures. Finally, the significant correlation between the estimated and observed MDS-UPDRS total scores shows that our framework is capable of identifying relevant features useful for disease severity assessment.

D. Selected Features for Diagnostic Categorization and Disease Severity Estimation

In the experiments, a small number of features were consistently and reliably selected across subjects and cross-validation iterations. For disease group classification, two dexterity features (one measures the variation in the number of touch events during a "right" tap, quantifying the variability of fine finger movement, and the other measures the time spent on screen during taps), two rest tremor features (one measures tremor frequency and the other measures tremor power in 3-8 Hz band) were consistently selected across the four different cross-validations, suggesting their reliability in distinguishing PD participants from HCs.

For MDS-UPDRS estimation, a dexterity feature (which indicates the variation in the number of touch events during a "right" tap, quantifying the variability of fine finger movement), a dexterity feature (which measures the distribution of time spent on screen during a tap), a gait feature (which quantifies the variability between the execution of different step phases), and a rest tremor feature (which measures the ratio of power in 5-6 Hz to total, quantifying acceleration power) were consistently selected, suggesting their reliability in evaluating PD severity.

Although selected features came from balance, dexterity, postural tremor, and rest tremor modalities, our results suggested that dexterity features made the most contribution to disease estimation. In **Figure** 2 (c), 5 out of 12 selected features were dexterity-related, and their weights (which determines their contribution to outcome estimation) were higher than those of others. Rest tremor, postural tremor, and balance features had smaller, but significant, weights and accounted for 4, 2, and 1 of the 12 selected features, respectively.

In feature selection, there is a trade-off to be made between accuracy, interpretability, and model parsimony. When α increases, most of the accuracy statistics improve (see **Figure** 3 (a)). Meanwhile, as α increases, the number of selected feature decreases (see **Figure** 3 (b)). A large α would yield a parsimonious model consisting of a small number of features. The elastic-net approach and the Lasso yield similar classification and estimation results. Typically, the Lasso arbitrarily picks a feature from a cluster of highly correlated features. The elastic-net, on the other hand, retains feature clusters in the model (where features within each cluster have similar influence), thereby avoiding making random within-cluster feature selection. This property is particularly attractive when investigators are interested in extracting several clusters of features across modalities, and in examining how consistent they are across different models.

V. Discussion

In this article, we introduced a machine-learning framework for conducting automated analyses of PD symptoms using smartphone data. Our proposed framework yielded accurate and meaningful results. Under this framework, we identified a reliable and PD-specific feature profile among individuals, using a dataset consisting of 437 features across six modalities measured on 17 unevenly sampled non-contiguous days during a period of up to six months. For PD classification, we obtained accuracy, specificity, and sensitivity statistics of 0.972, 0.971, and 0.973, respectively; for PD severity assessment, our estimated MDS-UPDRS total scores were significantly correlated with the observed scores ($r = 0.72, p < 0.0001$ using both PD and HC data and $r = 0.54, p < 0.001$ using only PD data) and yielded small RMSEs (16.58 using both PD and HC data and 17.19 using only PD data). The selected features were consistent with previous reports, where several distinct symptom domains, such as dexterity and rest tremor, present different patterns between PD participants and controls and are associated with disease outcomes [11], [13]–[24], [35]. Caution is warranted given the limited sample size; our analyses, however, suggest that the introduced model and the identified features are promising to assess PD in out-of-sample individuals. Future studies with even larger sample sizes should independently verify the extent to which these features are optimal to assess PD.

We have introduced and used a two-step feature selection procedure in the proposed machine-learning framework. Compared to a one-step feature procedure (*i.e.* using regularization alone), a two-step feature selection not only improves, but also balances, specificity and sensitivity (see **Table** 2).

For binary classification, we used a feature-wise t-test (which required a Gaussian assumption) during the first step feature selection (see **Section** III-B). Naturally, one would ask, by doing so, was it likely to overlook biomarkers that were not normally distributed but may be useful for identifying patients? To evaluate this possibility, we performed an additional analysis considering a U-test (or the Mann–Whitney test, which did not require a Gaussian assumption) [36] during the first step feature selection, while keeping everything else the same. Our results showed that it yielded slightly worse classification performance (Accuracy = 0.96 or 72/75 and AUC=0.99) than it of the proposed method (Accuracy = 0.973 or 73/75 and AUC = 0.991). A possible explanation can be made using a *bias-variance trade-off* argument. We achieved better results using a t-test than using a U-test because (1) when the (Gaussian) assumption held, the t-test may be more powerful than (at least as powerful as) the U-test; (2) in the case where the distributional assumption did not hold, we were trading variance with bias by using the t-test instead of the U-test. More precisely, by doing so we were more likely to incur wrong hypothesis testing result during the first step feature selection - although it was, in part, protected by using regularization in the second step feature selection - but the overall power was higher. To sum up, combining a t-test and regularization achieved higher bias (regarding training data, namely, it may not capture some regularity of the training data) and lower variance (in out-of-sample testing), and combining a U-test and regularization had lower bias (regarding training data, namely, it may better capture the regularities of the training data) and higher variance (in out-of-sample testing). As prediction was the main goal of automated disease assessment in mobile health, we chose to consider the combination of a t-test and regularization during the two-step feature selection procedure in disease classification.

We used the first moment to summarize features. The reasons for doing this were twofold. First, it is a simple, but effective, de-noising approach, evidenced by the experimental results in **Section** IV. The rational is that the mean can capture intrinsic properties of features for disease assessment problems, such as PD/HC classification, as some mean features differ significantly between PD participants and HCs. Second, it is computationally efficient. Compared to using the raw data directly, the mean approach incurs far less computing time, and yields improved estimation accuracy (see **Table** 2).

We would like to note several limitations of our framework. First, it only captures linearly associated behavioral features. Future work using non-linear parametric approaches may be useful to uncover the non-linear feature architecture related to PD. Second, it may overestimate disease severity when responses are zero-inflated (*i.e.* outcomes contain many entries equal or close to zero while other entries are very different from zero). Future work using Zero-Inflated Poisson (ZIP) model, and mixed-effect models may be useful to address this issue. Finally, our framework does not capture the features' temporal dynamics. Future work considering sliding-window analysis, generalized estimation equation (GEE) and generalized linear models (GLMs) with repeated measures [37], [38], and functional principal component analysis (fPCA) [39], [40], may be useful to unveil semi-continuous (windowed) and continuous temporal dynamic of the disease profile.

VI. Conclusions

Throughout this article, we have introduced and demonstrated an automated disease assessment framework that has the potential for remote PD classification and PD severity estimation at home using smartphones. Using this framework, we showed the presence of a disease-specific feature profile across multiple behavioral modalities.

Selected features reflected individually specific traits that were unique and generally reliable within subjects across 17 unevenly sampled non-contiguous days during a period of up to six months. Our results showed that the proposed framework is possible, with relatively high accuracy, to identify PD participants solely using sensor feature data from their remote, smartphone-based behavioral measurements. In addition, data analyses using this framework revealed that individual variability in extracted features (in particular, dexterity, rest tremor, and gait) were informative for the continuous disease severity estimation in an independent group of participants.

In conclusion, we proposed a machine-learning framework for automated disease assessment and provided preliminary evidence for a PD-specific behavioral architecture, in this case the extracted features, that may be associated with PD. Extensive data analyses suggest that this framework has the potential for identifying behavioral signatures to advance automated and remote assessment of PD.

Acknowledgment

The authors wish to thank Andreas U. Monsch of University of Basel, Switzerland (the PI on the Healthy Control study), and the study participants. The authors would also like to thank Timothy Kilchenmann, Detlef Wolf, Alf Scotland, Jens Schjodt-Eriksen, Wei-Yi Cheng, Ignacio Fernandez-Garcia, Juliane Siebourg-Polster, Liping Jin, Lynne Verselis, Frank Boess, Ronald B. Postuma, Anirvan Ghosh, Thomas Kremer, and Christian Czech, of Roche. Additionally, the authors would like to thank Jay Soto, Martin Koller, and Michael Grundman, of Prothena. Finally, the authors wish to thank the colleagues involved in developing and testing the app, deploying the devices, working with the hospitals, and processing the data, including Nicholas Heinemeier, Alessandro Barbato, and Jan Beckmann.

References

[1] L. M. De Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006.

[2] J. Prince and M. De Vos, "A deep learning framework for the remote detection of Parkinson's disease using smart-phone sensor data," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, 2018, pp. 3144–3147.

[3] A. Zhan, S. Mohan, C. Tarolli *et al.*, "Using smartphones and machine learning to quantify Parkinson disease severity: The mobile Parkinson disease score," *JAMA neurology*, vol. 75, no. 7, pp. 876–880, 2018.

[4] J. Prince, S. Arora, and M. De Vos, "Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes," *Physiological measurement*, vol. 39, no. 4, p. 044005, 2018.

[5] B. M. Bot, C. Suver, E. C. Neto *et al.*, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Scientific data*, vol. 3, p. 160011, 2016.

[6] A. Zhan, M. A. Little, D. A. Harris *et al.*, "High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection," *arXiv preprint arXiv:1601.00960*, 2016.

[7] E. C. Neto, T. M. Perumal, A. Pratap *et al.*, "On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: the mpower case study," *arXiv preprint arXiv:1706.09574*, 2017.

[8] S. Arora, V. Venkataraman, A. Zhan *et al.*, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study," *Parkinsonism & related disorders*, vol. 21, no. 6, pp. 650–653, 2015.

[9] C. Y. Lee, S. J. Kang, S.-K. Hong *et al.*, "A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease," *PloS one*, vol. 11, no. 7, p. e0158852, 2016.

[10] T. Arroyo-Gallego, M. J. Ledesma-Carbayo, Á. Sánchez-Ferro *et al.*, "Detection of motor impairment in Parkinson's disease via mobile touchscreen typing," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 1994–2002, 2017.

[11] P. Kassavetis, T. A. Saifee, G. Roussos *et al.*, "Developing a tool for remote digital assessment of Parkinson's disease," *Movement Disorders Clinical Practice*, vol. 3, no. 1, pp. 59–64, 2016.

[12] B. P. Printy, L. M. Renken, J. P. Herrmann *et al.*, "Smartphone application for classification of motor impairment severity in Parkinson's disease," in *Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE*, Chicago, IL, 2014, pp. 2686–2689.

[13] A. Benba, A. Jilbab, and A. Hammouch, "Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines," *International Journal on Electrical Engineering and Informatics*, vol. 7, no. 2, pp. 297–307, 2015.

[14] T. Kapoor and R. Sharma, "Parkinson's disease diagnosis using Mel-frequency cepstral coefficients and vector quantization," *International Journal of Computer Applications*, vol. 14, no. 3, pp. 43–46, 2011.

[15] J. Timmer, C. Gantert, G. Deuschl *et al.*, "Characteristics of hand tremor time series," *Biological cybernetics*, vol. 70, no. 1, pp. 75–80, 1993.

[16] G. Deuschl, J. Raethjen, R. Baron *et al.*, "The pathophysiology of Parkinsonian tremor: a review," *Journal of neurology*, vol. 247, no. 5, pp. V33–V48, 2000.

[17] G. Deuschl, J. Raethjen, M. Lindemann *et al.*, "The pathophysiology of tremor," *Muscle & nerve*, vol. 24, no. 6, pp. 716–735, 2001.

[18] D. E. Vaillancourt and K. M. Newell, "The dynamics of resting and postural tremor in Parkinson's disease," *Clinical Neurophysiology*, vol. 111, no. 11, pp. 2046–2056, 2000.

[19] A. Salarian, H. Russmann, C. Wider *et al.*, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pp. 313–322, 2007.

[20] A. L. Taylor Tavares, G. S. Jefferis, M. Koop *et al.*, "Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation," *Movement disorders*, vol. 20, no. 10, pp. 1286–1298, 2005.

[21] R. E. Mayagoitia, A. V. Nene, and P. H. Veltink, "Accelerometer and rate gyroscope measurement of kinematics: an inexpensive alternative to optical motion analysis systems," *Journal of biomechanics*, vol. 35, no. 4, pp. 537–542, 2002.

[22] F. B. Horak and M. Mancini, "Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors," *Movement Disorders*, vol. 28, no. 11, pp. 1544–1551, 2013.

[23] M. Mancini, P. Carlson-Kuhta, C. Zampieri *et al.*, "Postural sway as a marker of progression in Parkinson's disease: a pilot longitudinal study," *Gait & posture*, vol. 36, no. 3, pp. 471–476, 2012.

[24] W.-Y. Cheng, A. Scotland, F. Lipsmeier *et al.*, "Human activity recognition from sensor-based large-scale continuous monitoring of Parkinson's disease patients," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Philadelphia, PA, 2017, pp. 249–250.

[25] F. Lipsmeier, K. Taylor, T. Kilchenmann *et al.*, "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial," *Movement Disorders*, vol. 33, no. 8, pp. 1287–1297, 2018.

[26] Z. S. Nasreddine, N. A. Phillips, V. Bédirian *et al.*, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[27] C. G. Goetz, B. C. Tilley, S. R. Shaftman *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[28] R. Bhidayasiri and P. Martinez-Martin, "Clinical assessments in Parkinson's disease: scales and monitoring," in *International review of neurobiology*. Elsevier, 2017, vol. 132, pp. 129–182.

[29] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[30] D. B. Rubin, *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004.

[31] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[32] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Doklady Akademii Nauk SSSR*, vol. 151, pp. 501–504, 1963.

[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[34] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, pp. 1–20, 2010.

[35] A. Salarian, C. Zampieri, F. B. Horak *et al.*, "Analyzing 180° turns using an inertial system reveals early signs of progression of Parkinson's disease," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, 2009, pp. 224–227.

[36] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[37] S. L. Zeger, K.-Y. Liang, and P. S. Albert, "Models for longitudinal data: a generalized estimating equation approach," *Biometrics*, vol. 44, no. 4, pp. 1049–1060, 1988.

[38] L. H. Moulton and S. L. Zeger, "Analyzing repeated measures on generalized linear models via the bootstrap," *Biometrics*, vol. 45, no. 2, pp. 381–394, 1989.

[39] F. Yao, H.-G. Müller, J.-L. Wang *et al.*, "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, vol. 33, no. 6, pp. 2873–2903, 2005.

[40] G. He, H. Müller, and J. Wang, "Extending correlation and regression from multivariate to functional data," in *Asymptotics in statistics and probability*, M. Puri, Ed. VSP, Zeist, 2000, pp. 197–210.