Exploring test retest reliability and longitudinal stability of digital biomarkers for Parkinson's disease in the m-Power dataset

Mehran Sahandi Far[1,2], Simon B. Eickhoff [1,2], María Goñi [1,2], Juergen Dukart [1,2]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
[2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany

Introduction

Standard in-clinic assessments such as UPDRS are :
- Costly
- limited with respect to observation frequency
- influenced by inter-rater variability

Digital biomarkers (DB) as captured using sensors embedded in modern smart devices are a promising technology for home-based symptom monitoring in Parkinson's disease (PD).

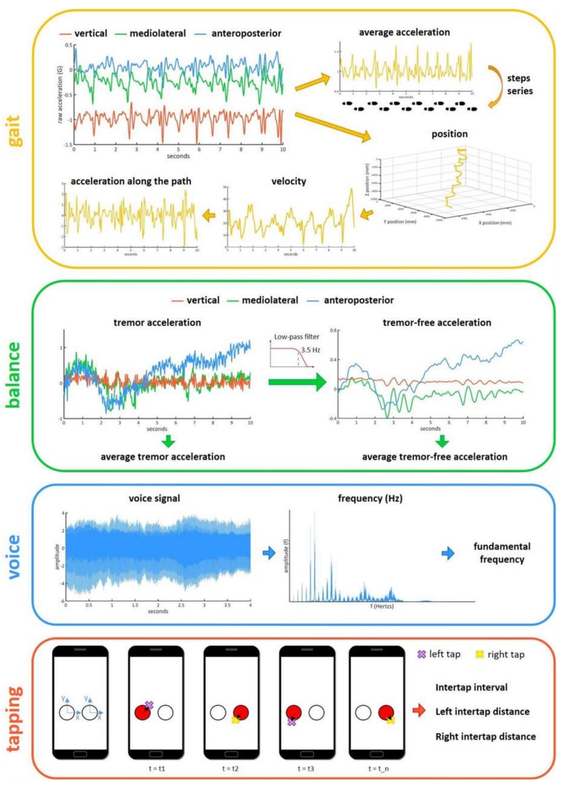
DB being collected frequently over a long period of time can provide :
- objective,
- ecologically valid
- better understanding of the inter- and intra-individual variability in disease manifestation in daily life.

DB measures are prone to large variation caused by technical and procedural differences such as:
- placement/orientation,
- recording frequency
- environmental and individual variation, i.e. due to motivation, medication or other aspects
- impact of learning
- users' familiarity with technology

No PD studies systematically **evaluated** the **test-retest reliability** and longitudinal **sensitivity** of DB in **a fully unsupervised and self-administered** PD longitudinal setting.

Methods

Feature extraction
An overall of 773 features related to gait (N=423), balance (N=183), finger dexterity (N=43), and speech impairment (N=124).



Statistical analysis

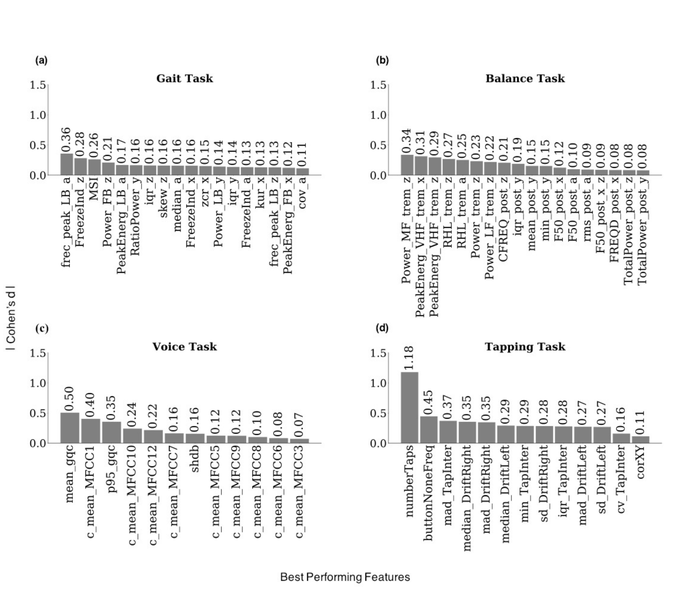
For features to be considered usable for biomarker :
- sensitivity to disease symptoms,
- good test-retest reliability
- robustness against the effects of learning and other longitudinal confounds.

Statistical analysis
Mann-Whitney U test were used to identify all features that significantly differ between PD and HC at the first administration (baseline) ($p<0.05$). Effect sizes Cohen's d were computed for these features to provide an estimate of the magnitude of differentiation between PD and HC.

Intraclass Correlation Coefficients (ICC, type 1-1) were used to determine the test-retest reliability of features showing a significant differentiation between PD and HC.

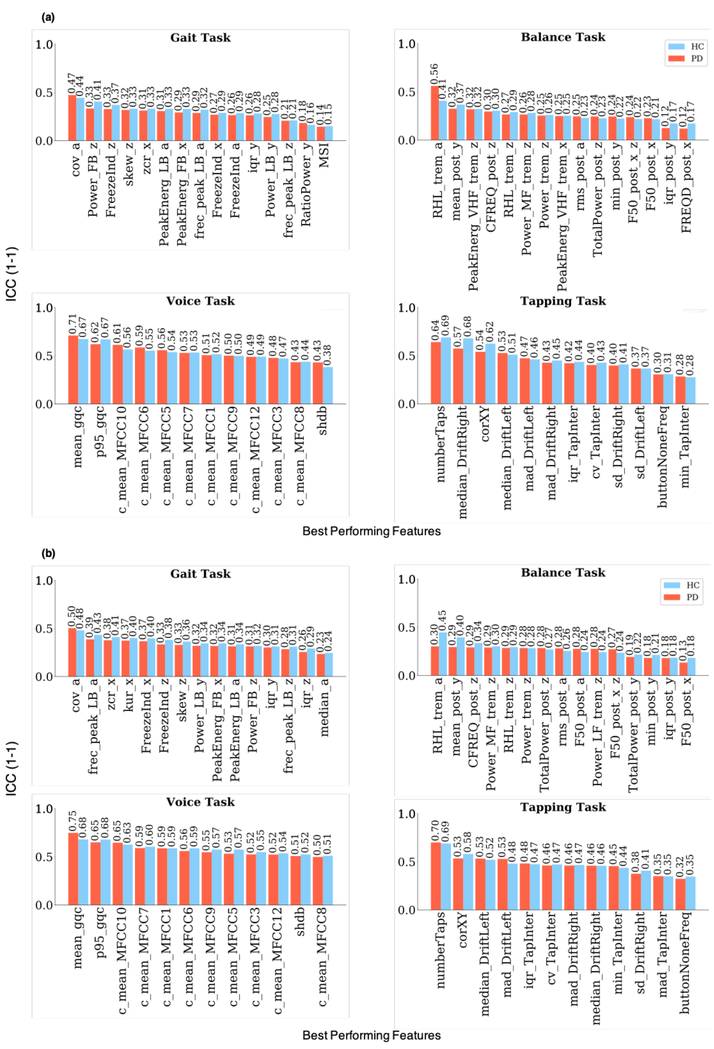
The impact of PD medication by computing rm-ANOVAs in the PD group with the within-subject factor medication (i.e. before, after, and at best).

Results

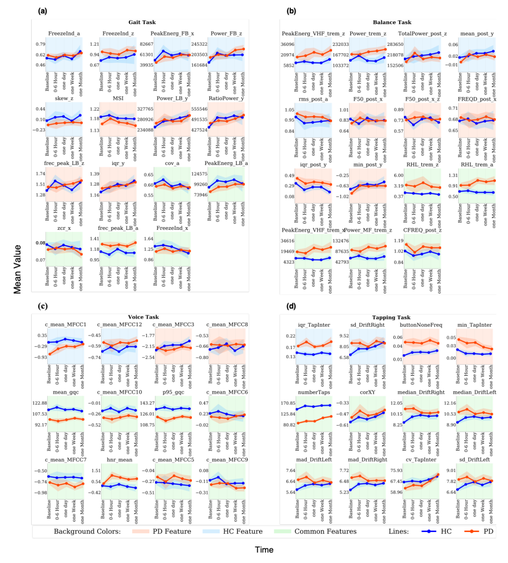


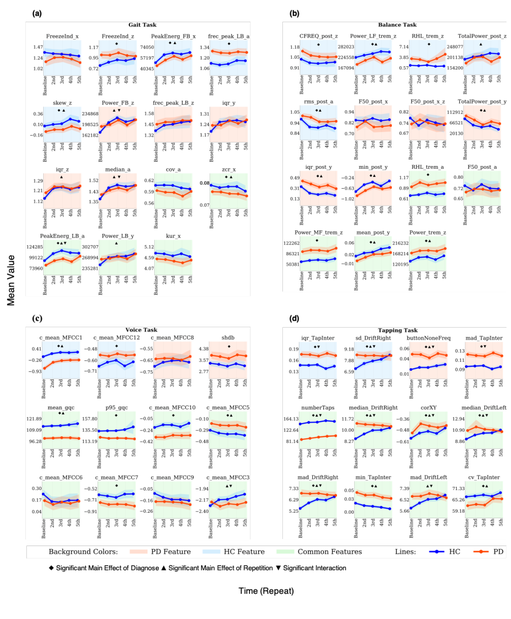
we performed group comparisons for all computed features for gait, balance, voice and tapping tasks. For gait 66 out of 423, for balance 59 out of 183, for voice 60 out of 124 and for tapping 25 out of 43 features differed significantly ($p<0.05$) between PD and HC at baseline with small (gait and balance) to medium effect sizes for gait, balance and voice and small to large effect sizes for the tapping task

A significant main effect of diagnosis across all time points was observed for 6 out of 15 gait features, 11 out of 15 balance features, 8 out of 12 voice features and 11 out of 12 tapping features. A significant effect of repetition was found for 8 out of 15 gait features, 8 out of 15 balance features, 4 out of 12 voice features 10 out of 12 tapping features. A significant diagnosis-by-repetition interaction effect was identified for 3 out of 15 gait features, 0 out of 15 balance features, 3 out of 12 voice features, and 9 out of 12 tapping features.



ICC analyses revealed poor to good test-retest reliability for the most reliable features from the gait and balance tasks and good to excellent reliability for features from voice and tapping tasks





Discussion

1) Only very few features had medium to large effects sizes for differentiation between PD and HC. For all tasks, a substantial percentage of features displayed significant longitudinal alterations.

2) Overall, tapping and voice tasks revealed a better performance compared to gait and balance tasks with respect to test-retest reliability and observed effect sizes.

3) Both, balance and gait tasks displayed consistently poor test-retest reliabilities as well as low effect sizes for differentiation between PD and HC questioning their usability for home-based applications.

4) Most features showed a drop in test-retest reliability with longer periods of time. This may potentially reflect a consequence of the repetition effects and the group-by-repetition interactions observed in the analyses of variance for a substantial proportion of the features.

1) alterations in PD relative to HC reflect impairment, movement of a feature state towards PD is likely to reflect worsening either due to reduced motivation, disease progression or other similar factors. In contrast, movements towards HC is likely to reflect improvement and is therewith compatible with a learning effect.
2) We find a mixture of both effects for most tasks suggesting the presence of both aspects in DB longitudinal data.
3) These observations are also in line with previous studies showing that training may reduce motor impairment in PD
4) These findings may point to a differential change in motivation across groups. Whilst differential learning has been previously reported [34,37,38], the differential change in motivation is an important novel aspect to consider when comparing DB measures between PD patients and HC.

5) **our findings clearly demonstrate the need for further optimization of DB tasks as well for introducing careful monitoring and quality control procedures to enable integration of DB measures into clinically relevant applications.**

References:Alexander, L.M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., et al. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. Sci. Data 4, 170181.