Machine learning for automated EEG-based biomarkers of cognitive impairment during Deep Brain Stimulation screening in patients with Parkinson's Disease

Geraedts, V.J.; Koch, M.; Contarino, M.F.; Middelkoop, H.A.M.; Wang, H.; Hilten, J.J. van; ... ; Tannemaat, M.R.

Contents lists available at ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

Machine learning for automated EEG-based biomarkers of cognitive impairment during Deep Brain Stimulation screening in patients with Parkinson's Disease

V.J. Geraedts [a,b,*], M. Koch [c], M.F. Contarino [a,d], H.A.M. Middelkoop [a], H. Wang [c], J.J. van Hilten [a], T.H.W. Bäck [c], M.R. Tannemaat [a]

[a] *Leiden University Medical Centre, Department of Neurology, the Netherlands*
[b] *Leiden University Medical Centre, Department of Epidemiology, the Netherlands*
[c] *Leiden Institute of Advanced Computer Science, the Netherlands*
[d] *Haga Teaching Hospital, Department of Neurology, the Netherlands*

ARTICLE INFO

Article history:
Accepted 12 January 2021
Available online 24 February 2021

Keywords:
Machine learning
Quantitative EEG
Parkinson's Disease
Deep Brain Stimulation
Cognition

HIGHLIGHTS

- A fully automated EEG-based machine learning pipeline was applied to DBS candidates with Parkinson's Disease.
- The model differentiates good versus poor cognitive function with high accuracy.
- Automatically extracted EEG biomarkers may have utility during the DBS screening.

ABSTRACT

Objective: A downside of Deep Brain Stimulation (DBS) for Parkinson's Disease (PD) is that cognitive function may deteriorate postoperatively. Electroencephalography (EEG) was explored as biomarker of cognition using a Machine Learning (ML) pipeline.

Methods: A fully automated ML pipeline was applied to 112 PD patients, taking EEG time-series as input and predicted class-labels as output. The most extreme cognitive scores were selected for class differentiation, i.e. best vs. worst cognitive performance (n = 20 per group). 16,674 features were extracted per patient; feature-selection was performed using a Boruta algorithm. A random forest classifier was modelled; 10-fold cross-validation with Bayesian optimization was performed to ensure generalizability. The predicted class-probabilities of the entire cohort were compared to actual cognitive performance.

Results: Both groups were differentiated with a mean accuracy of 0.92; using only occipital peak frequency yielded an accuracy of 0.67. Class-probabilities and actual cognitive performance were negatively linearly correlated ($\beta = -0.23$ (95% confidence interval (-0.29, -0.18))).

Conclusions: Particularly high accuracies were achieved using a compound of automatically extracted EEG biomarkers to classify PD patients according to cognition, rather than a single spectral EEG feature.

Significance: Automated EEG assessment may have utility for cognitive profiling of PD patients during the DBS screening.

1. Introduction

Parkinson's Disease (PD) is the fastest growing neurological disorder worldwide, with both characteristic motor and non-motor symptoms. Patients who develop motor complications may be eligible for Deep Brain Stimulation (DBS), an invasive surgical intervention which is highly effective in relieving motor complications and improves quality of life (Ahlskog and Muenter, 2001; Deuschl and Agid, 2013). Despite good effects on motor functioning and substantial relief of motor complications refractory to oral medication (Deuschl and Agid, 2013; Okun et al., 2012), DBS does not improve cognitive symptoms and some deterioration can be observed in cognitive domains (Contarino et al., 2007; Weaver et al., 2009) and neuropsychi-

* Corresponding author at: Departments of Neurology and Epidemiology, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, the Netherlands.

E-mail address: v.j.geraedts@lumc.nl (V.J. Geraedts).

https://doi.org/10.1016/j.clinph.2021.01.021
1388-2457/© 2021 International Federation of Clinical Neurophysiology. Published by Elsevier B.V.

atric functioning after surgery (Drapier et al., 2006; Smeding et al., 2011). The screening process for DBS therefore entails an extensive evaluation of cognitive and neuropsychiatric functioning to rule out severe impairment prior to surgery, in order to determine DBS eligibility (Geraedts et al., 2019; Lang et al., 2006). However, accurate evaluations of cognition are limited by factors such as intellectual status (Duncan, 1993), while performance tasks may be subject to misinterpretation due to e.g. motor impairment, fatigue, mood disorder, stress, and personal motivation, which may render results less valid (Duckworth et al., 2011; Duckworth and Yeager, 2015). In addition, neuropsychological screening is time-consuming and stressful for patients. Consequently, there is a need for new biomarkers to complement current neuropsychological assessments of cognition.

A candidate instrument for such complementary assessments is quantitative Electroencephalography (qEEG), which can measure brain activity directly and non-invasively. The utility of qEEG to aid during assessment of cognitive impairment, and even predict cognitive deterioration has been previously established in the general PD population (Geraedts et al., 2018a). Particularly spectral features reflecting EEG slowing are related to cognitive deterioration, although recent advances in EEG processing have demonstrated an association of cognitive impairment with connectivity and network dysfunction in cross-sectional studies as well (Chaturvedi et al., 2019; Geraedts et al., 2018b; Utianski et al., 2016). However, these latter metrics have been sparsely studied in comparison to spectral analyses (Geraedts et al., 2018a). An extensive evaluation across the numerous possibilities of EEG metrics beyond spectral powers, to determine which metrics have the highest potential for reflecting PD symptoms, is lacking.

A limitation of qEEG analyses is the laborious amount of pre-processing, and particularly, the arbitrary selection of features to include during the final modelling. Traditionally, features from time series are manually selected and computed, which is time-consuming and requires expert knowledge and is therefore difficult to translate to clinical practice. A machine learning (ML) approach may overcome these limitations by providing output, such as a classification of cognitive status, without predefined data-extraction or modelling (Bonanni, 2019). Preliminary ML results on determining levels of cognitive severity demonstrated high performance scores, although limited to predetermined (spectral) features only. These models still require a large degree of pre-processing and manual feature-extraction (Betrouni et al., 2019). Ideally, the ML approach is extended to a fully automated ML pipeline, deemed a 'sequence of data processing components' (Geron, 2017). Within a ML pipeline, the EEG time series are delivered as input, after which an automated algorithm extracts a large number of features, selects those features which are needed to create a representative EEG profile, and learns and optimizes a ML model, without any intervention in between. Such a pipeline limits the necessity of making arbitrary choices, makes the entire process more efficient, and increases the likelihood of identifying novel biomarkers.

Given the need for complementary objective screening instruments to evaluate cognition during the DBS screening, the aim of our study was to evaluate the utility of a qEEG ML pipeline for determining cognitive status in these patients. To this end, the most 'extreme' DBS candidates were selected to build a supervised learning model, i.e. best vs. worst cognitive scores after a comprehensive neuropsychological test battery. The model could then be applied to evaluate the remaining DBS candidates, during which the association between ML-predictions and the actual levels of cognitive function could be studied.

2. Methods

2.1. Study participants

All consecutive patients who underwent preoperative screenings for DBS at the Leiden University Medical Center (LUMC) between September 2015 and June 2019 were included in the study. All patients fulfilled the criteria for clinically established PD (Postuma et al., 2015). The study was approved by the local medical ethics committee and all patients gave written informed consent.

2.2. EEG acquisition, pre-processing and analysis

EEG acquisition and pre-processing has been described elsewhere (Geraedts et al., 2018b). Eyes-closed resting-state recordings were made with 21 Ag/AgCl EEG electrodes according to standard 10–20 positions. Patients used their medication according to their individual schedules (i.e. 'ON'); dyskinesias were not observed. Data were re-referenced towards a source derivation approaching the surface Laplacian derivation (Hjorth, 1980) to amplify spatial resolution (Burle et al., 2015). After visual confirmation of artefact-free signals, five consecutive non-overlapping 4096-point (8.192 seconds) epochs were selected for offline analysis in American Standard Code for Information Interchange (ASCII) format. Recordings with less than five epochs were excluded from analyses. Brainwave software was used for computation of clinically used peak frequencies (BrainWave version 0.9.152.12.26, C. J. Stam; available at http://home.kpn.nl/stam7883/brainwave.html).

2.3. Group composition

From the comprehensive neuropsychological evaluations, six neuropsychological domains were identified according to the Diagnostic and Statistical Manual of mental disorders (5th edition, DSM-V).(American Psychiatric Association, 2013) According to DSM-V consensus guidelines, the following cognitive tests were selected for each domain: (1) 'Learning and Memory': Cambridge Cognitive Examination (CAMCOG) memory section (Huppert et al., 1995), Rey Auditory Verbal Learning Test (RAVLT) (Vakil and Blachstein, 1993), and Wechsler Memory Scale (WMS) (Wechsler, 1945); (2) 'Executive Functioning': CAMCOG abstract reasoning, Digit Cancellation Test (DCT) (Dekker et al., 2007), digit span (Richardson, 2007), Word-colour Stroop Test (Stroop) 3 (Scarpina and Tagini, 2017), Trail Making Test (TMT) B (Tombaugh, 2004); (3) 'Psychomotor speed': Stroop 1 and 2, and TMT A; (4) 'Language': CAMCOG language section and verbal fluency; (5) 'Perceptive-motoric functioning': CAMCOG perception and CAMCOG praxis, and (6) 'Neuropsychiatric status': Becks Depression Inventory (BDI) (Beck et al., 1996) and Hospital Anxiety and Depression Scale (HADS) A-D (Zigmond and Snaith, 1983). All individual test-scores were standardised (Z-transformed) and averaged per domain for direct comparability. In case of missing data, an average of the remaining test-scores within the pertaining domain was used rather than imputing data, as long as ≥2 test-scores remained per domain (except for the domain 'Language' which contains only two tests and for which no data was imputed). A composite Z-score was derived from averaging all domains, if data from ≥4 domains were available. Higher Z-scores indicate better cognitive functioning. From the entire dataset, the most extreme patients in terms of cognitive performance were selected: either the highest cognitive composite scores (high-COG, n = 20) or the lowest scores (low-COG, n = 20). All other patients were classified as 'intermediate cognitive performance (int-COG). Given the nature of the cohort (i.e. DBS candidates who had already underwent a clinical pre-screening) (Geraedts et al., 2019), it was deemed unlikely that a sufficient number of patients would fulfil the criteria for either PD Dementia (PDD) or Mild Cognitive Impairment (MCI) and these classes were therefore deemed unsuitable to use for classification purposes.

Secondary outcomes included: motor function (Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) part III (range 0–132)) (Goetz et al., 2008), and non-dopaminergic functioning (SEverity of Non-dopaminergic Symptoms in Parkinson's Disease (SENS-PD) scale (range 0–54)) (van der Heeden et al., 2016), and level II criteria for PD-MCI (Litvan et al., 2012).

2.4. ML pipeline

A previously reported ML pipeline approach was used for time series classification purposes (Koch and Bäck, 2018; Koch et al., 2018). Originally developed and applied in the automotive industry to classify time series originating from vehicle-data (i.e. predicting damaged parts after a low-speed crash (Koch and Bäck, 2018; Koch et al., 2018)), the approach was further applied to time series originating from EEGs, particularly to evaluate different ML approaches for classification of PD patients according to their cognitive performance (Koch et al., 2019). The resulting ML pipeline consists of four phases: (1) feature-extraction, (2) feature-selection, (3) training of a classifier, and (4) hyperparameter optimization. All four steps are completely automated, with the EEG time series as input and the class-labels (i.e. high-COG or low-COG) as output. The library 'Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests' (tsfresh) was used to extract features from the time series (Christ et al., 2018, 2016), resulting in 16,674 features per EEG (794 comprehensive features for each of the 21 time series) (Kursa and Rudnicki, 2010). Feature selection was performed using the Boruta algorithm, by testing the variable importance (VIMP) of each feature against that of 'shadow features', which are created by random shuffling of the real features. The VIMP of shadow and real features are obtained from a random forest model trained thereon. A real feature would be selected if its VIMP frequently dominates the maximal VIMP of shadow features, in multiple independent trials (Kursa and Rudnicki, 2010). After feature-selection, this feature set is used to train a Random Forest Classifier (RFC). A RFC is an ensemble of decision trees; the resulting decision is the majority vote from all decision trees (Hastie et al., 2009). The hyperparameters of the RFC, such as the number of decision trees and their individual tree depths, are optimized with a variant of Bayesian Optimization technique called Mixed Integer Parallel Efficient Global Optimization (MIP-EGO) (Wang et al., 2018, 2017) for mixed-integer categorical search spaces (Yang et al., 2019). To ensure generalizability of the RFC, a cross-validation procedure was adopted: the data is randomly split into 10 folds, after which training was performed on 9 folds and tested on the remaining fold. This process was repeated until each fold has served as test set; the average of all test scores of the computations represents the final score. A secondary assessment of interval validity was based on a combination of cross-validation and split-sample validation: cross-validated model-training based on 50% of the data and validated on the remaining sample. This approach was repeated for 60–90% of the data used for model-building with the remaining sample used for internal validation purposes, although it should be noted that cross-validation is superior to split-sample validation to assess internal validity especially for small sample sizes (Steyerberg, 2018). A detailed description of the applied ML Pipeline is published elsewhere (Koch et al., 2019). Since all four steps are fully automated, no arbitrary choices on feature-extraction or feature-selection were made during the model-building-process.

2.5. Application of the pipeline to EEG data

Both occipital and global peak frequencies, routinely used for clinical purposes, were used as standard-features. All five epochs were averaged per patient, in order to obtain more robust time series and to limit intra-individual variability (Koch et al., 2019). The features from each individual computation-run were selected and combined. The resulting model with the combined features was evaluated for model performance. A comparison was drawn between a model using only the occipital peak frequency as a single classifying feature and the ML Pipeline using a combination of the routinely-used peak frequency and the automatically extracted features from the EEG time series.

The final selected model with the best-classifying performance was then applied to the unclassified patients (i.e. those with 'intermediate' cognitive performance scores) and the predicted probabilities of being classified as low-COG were calculated for all patients. A linear regression model was fitted with these predicted probabilities as an outcome, and the composite global cognitive score subdivided into three splines in accordance with the original cognitive classification as independent variables.

2.6. Statistical analysis

Demographic, clinical, and neuropsychological variables, as well as electrophysiological spectral features, were compared between the high-COG and low-COG groups using Student T-tests if normally distributed, and Mann-Whitney U tests if not-normally distributed in case of continuous variables, and Pearson's χ^2 Tests in case of categorical data. The ML Pipeline, as well as a model using only occipital peak frequency as classifying feature, was evaluated using accuracy, sensitivity, and specificity metrics. The features included in the ML pipeline were compared using General Linear Models, both crude and corrected for age, disease duration, and sex.

Missing values, other than cognitive performance scores, were imputed using multiple imputation with five iterations in case of $\leq$15% missing data.

All analyses were performed using IBM Statistical Package for the Social Sciences (SPSS) 25 Software (SPSS inc., Chicago, Illinois, USA).

2.7. Data availability

Anonymized data may be shared upon request.

3. Results

3.1. Patient characteristics

A total of 112 patients were included. Patients classified as high-COG were younger, and with a younger age-at-onset than low-COG patients. Non-dopaminergic disease severity, as well as motor functioning during 'ON' was better in high-COG patients, whereas motor functioning during 'OFF' did not differ (see Table 1). Composite cognitive Z scores were inherently different between the high-COG and low-COG groups with approximately 1.5 standard deviations (SD) difference (mean (SD) 0.78 (0.57) vs. −0.78 (0.54), respectively). High-COG patients had similarly better scores for the domains 'Learning and Memory', 'Perceptive-motoric functioning', 'Executive functioning', and 'Language'. Strikingly, scores for the domains 'Neuropsychiatric functioning' and 'Psychomotoric

Table 1
Demographic and clinical characteristics.

	High-COG	Low-COG	P [*]	Int-COG
N	20	20		72
Age [a]	59.5 (54.6–66.4)	67.8 (60.1–72.1)	0.004	63.5 (57.7–68.0)
Age at onset [b]	48.2 (9.3)	55.4 (9.6)	0.023	51.1 (10.7)
Disease duration [b]	11.2 (4.5)	10.9 (5.1)	0.814	11.8 (8.0)
LED [a]	1151.50 (900.00–1287.50)	1097.25 (517.50–1519.13)	0.547	1150.00 (811.50–1463.50)
% Use of psychoactive medication (n) [c]	15 (3)	30 (6)	0.451	32 (23)
% Female (n) [c]	45 (9)	10 (2)	0.031	37.5 (27)
MDS-UPDRS III 'ON' [a]	18.5 (11–22.5)	23 (19–36)	0.012	20.5 (13.3–30)
MDS-UPDRS III 'OFF' [a]	46.5 (39.3–55.5)	48.5 (41–57)	0.718	44 (36–55)
SENS-PD [b]	9.2 (4.0)	15.3 (4.8)	<0.001	12.4 (4.8)
Z Psychomotoric speed [a]	−0.71 (−0.97 to −0.38)	0.55 (−0.27 to 1.30)	<0.001	−0.23 (−0.60 to 0.18)
Z Language [a]	0.88 (0.50–1.24)	−0.93 (−2.11 to −0.45)	<0.001	0.04 (−0.35 to 0.53)
Z Neuropsychiatric functioning [a]	−0.40 (−0.78 to 0.28)	0.16 (−0.39 to 0.41)	0.108	−0.12 (−0.42 to 0.37)
Z Executive functioning [a]	0.59 (0.28–0.74)	−0.71 (−1.64 to −0.35)	<0.001	0.08 (−0.23 to 0.40)
Z Perceptive-motoric functioning [a]	0.40 (0.40–0.76)	−1.35 (−1.61 to −0.63)	<0.001	0.40 (−0.06 to 0.76)
Z Learning and Memory [a]	0.92 (0.34–1.07)	−0.79 (−1.83 to −0.32)	<0.001	0.06 (−0.28 to 0.50)
Z Global Cognition [b]	0.78 (0.57)	−0.78 (0.54)	<0.001	0.09 (0.22)
% PD-MCI (≥2 domains ≤ −1.5 SD) (n)	0	30 (6)		0
% PD-MCI (≥2 domains (−1, −1.5) SD) (n)	0	15 (3)		3 (2)

Int-COG = all patients with intermediate cognitive scores.

LEDD: Levodopa Equivalent Dose; PD-MCI: Parkinson's Disease Mild Cognitive Impairment MDS-UPDRS III: Movement Disorders Society – Unified Parkinson's Disease Rating Scale III; SENS-PD: SEverity of Non-dopaminergic Symptoms in Parkinson's Disease.

[*] High-COG (20 patients with highest cognitive scores) vs. Low-COG (20 patients with lowest cognitive scores)

[a] Mann Whitney U tests (median (interquartile range)).

[b] Student T tests (mean (standard deviation)).

[c] Pearson $\chi 2$ tests.

speed' were lower for the high-COG patients than for the low-COG patients.

High-COG patients had spectrally faster EEGs than low-COG patients, demonstrated by particularly higher occipital peak frequencies (mean (SD) 9.0 (0.9) vs. 7.8 (1.4) Hz) and lower ratios of slow-over-fast relative powers ($(\delta + \theta)/(\alpha 1 + \alpha 2 + \beta)$) (median (interquartile range) 0.69 (0.49–0.86) vs. 1.21 (0.57–2.20) (Table 2 and Fig. 1).

Patients classified as int-COG had clinical, cognitive, and spectral scores situated between low-COG and high-COG scores, respectively.

3.2. ML pipeline performance

The accuracy (mean (SD)) of the average of all individual runs of the pipeline was 0.81 (0.01). After a secondary series of cross-validation runs incorporating all features from the individual runs, the extended model performance increased to 0.92 (0.02). Using only the occipital peak frequency as a classifying feature, the accuracy was lower: 0.67 (0.06) (see Table 3). The list of features (n = 13) selected by the ML pipeline included the clinically used 'occipital peak frequency'. No significant differences were found for the included features (see Supplementary Table S1), except for the occipital peak frequency, both in the analysis including crude differences and after correction for age, disease duration, and sex. All features were in a VIMP range of 4–15% (see Supplementary Table S1 and Supplementary Figure S1A-E for a complete overview). A combination of cross-validation and split-sample validation demonstrated good internal validity for all splits (see Supplementary Figure S2).

An additional model differentiating low-COG from int-COG yielded a mean accuracy of 0.80 (0.03); whereas an additional model differentiating int-COG from high-COG yielded a mean (SD) accuracy of 0.80 (0.02). As classes were relatively unbalanced, the sensitivity of the models was much lower (low-COG vs. int-COG: mean (SD) sensitivity 0.26 (0.16); int-COG vs. high-COG mean (SD) sensitivity: 0.24 (0.08)), and corresponding specificities were relatively high (low-COG vs. int-COG: mean (SD) specificity 0.95 (0.03); int-COG vs. high-COG mean (SD) 0.96 (0.04)) (see Supplementary Table S2).

3.3. Calibration

A scatterplot demonstrating the correlation between actual cognitive functioning and the predicted probability of being classified as low-COG is shown in Fig. 2, demonstrating a negative trend (i.e.

Table 2
EEG spectral characteristics.

	High-COG	Low-COG	P [*]	Int-COG
Occipital peak frequency [a]	9.0 (0.9)	7.8 (1.4)	0.003	8.4 (1.4)
Total peak frequency [a]	8.8 (0.8)	7.9 (1.4)	0.013	8.2 (1.1)
Relative δ power [b]	0.21 (0.18–0.27)	0.24 (0.17–0.39)	0.369	0.26 (0.20–0.35)
Relative θ power [b]	0.15 (0.11–0.20)	0.20 (0.13–0.31)	0.068	0.17 (0.12–0.26)
Relative $\alpha 1$ power [b]	0.23 (0.16–0.30)	0.16 (0.07–0.22)	0.024	0.14 (0.09–0.21)
Relative $\alpha 2$ power [b]	0.11 (0.09–0.17)	0.07 (0.06–0.11)	0.008	0.09 (0.06–0.13)
Relative β power [b]	0.19 (0.16–0.25)	0.16 (0.12–0.23)	0.327	0.19 (0.15–0.25)
Slowing ratio ($(\delta + \theta)/(\alpha 1 + \alpha 2 + \beta)$) [b]	0.69 (0.49–0.86)	1.21 (0.57–2.20)	0.026	1.07 (0.59–1.43)

High-COG (20 patients with highest cognitive scores) vs. Low-COG (20 patients with lowest cognitive scores).

Int-COG = all patients with intermediate cognitive scores.

[a] Student T-test (mean (standard deviation)).

[b] Mann Whitney U test (median (interquartile range)).

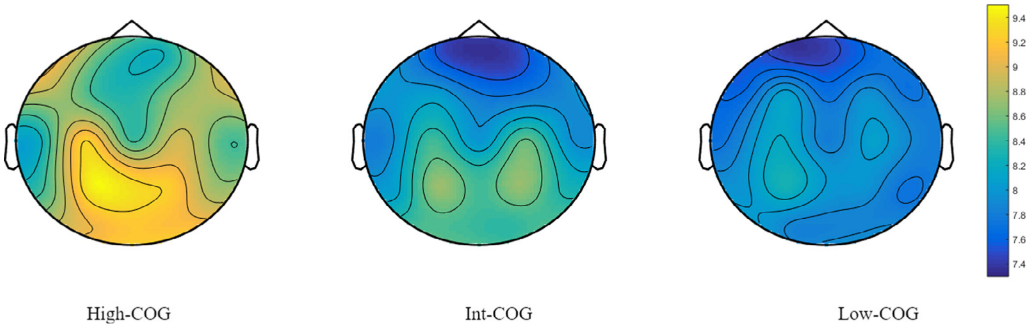

Fig. 1. Spectral plots (peak-frequency) per cognitive class. Peak frequencies were calculated in Hz. Patients with high cognitive performance scores have spectrally faster EEGs than patients with lower cognitive performance scores. Low-COG: lower cognitive performance scores; Int-COG: intermediate cognitive performance scores; High-COG: higher cognitive performance scores.

Table 3
Machine Learning model performances.

	Occipital peak frequency only	Mean of all individual cross-validation runs	All features from all cross-validation runs
Accuracy	0.67 (0.06)	0.81 (0.01)	0.92 (0.02)
Sensitivity	0.74 (0.09)	0.82 (0.04)	0.90 (0.04)
Specificity	0.59 (0.04)	0.83 (0.07)	0.94 (0.02)

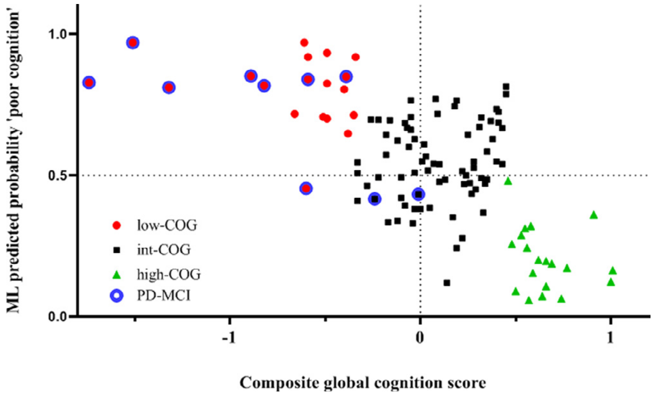

Fig. 2. Predicted probability of being classified 'low-COG' vs. actual cognitive performance. Low-COG: lower cognitive performance scores; Int-COG: intermediate cognitive performance scores; High-COG: higher cognitive performance scoers; PD-MCI: Parkinson's Disease – Mild Cognitive Impairment; ML: Machine Learning.

a lower probability of being classified as low-COG correlates to better cognition: β = −0.23 (95%CI −0.29, −0.18)). Both the high-COG and the low-COG groups contributed to this negative trend (spline-high-COG: β = −0.289 (95% CI −0.37, −0.20), spline-low-COG: β = −0.26 (95%CI −0.34, −0.17)), but the int-COG patients, who were not used during model-training, did not (spline-int-COG: β = 0.12 (95%CI −0.05, 0.30)).

4. Discussion

In this study, we show that DBS candidates with PD with either clinically determined 'good' or 'poor' cognition may be classified according to their cognitive function based on a fully automated EEG-assessment.

Contrary to previous studies which highlight singular, or few features to distinguish patients with different levels of cognitive impairment (Betrouni et al., 2019; Chaturvedi et al., 2019; Geraedts et al., 2018b; Klassen et al., 2011; Utianski et al., 2016), we showed that a compound of multiple EEG-biomarkers provides the highest accuracy in classifying patients.

Our final model performs slightly better than previously reported ML algorithms, which report accuracies between 74% (Chaturvedi et al., 2019) and 88%.(Betrouni et al., 2019) Betrouni and colleagues differentiated five groups of PD patients, with different levels of cognitive impairment using support vector machines (accuracy = 84%) and k-nearest neighbour models (88%) (Betrouni et al., 2019). Although different electrode-densities were used, analyses were limited to spectral features in an effort to prevent overfitting. As the dataset was subdivided into five different categories based on cognitive clusters, the two groups with worst cognitive function were smallest, containing respectively five and nine patients. In contrast, the results described above demonstrate the advantage of automated feature-extraction and simultaneous analysis to both increase the accuracy and limit the need for laborious pre-processing. Pragmatically, the use of spectral features to reflect EEG slowing is currently still easier to translate to routine clinical practice than applying a ML pipeline to new EEG data, although less accurate. Another study added connectivity metrics, i.e. Phase-Lag-Index (PLI) to spectral features resulting in 396 features (66 spectral- and 330 PLI features) (Chaturvedi et al., 2019). Although the reported accuracies were lower, PLI features discriminated better between PD patients with or without MCI (spectral features: Area-under-the-curve (AUC) = 0.64; PLI features: AUC = 0.74). Our model does not include between-channel connectivity metrics but rather focuses on synchronization patterns within one individual time series. The amount of computation runtime increases exponentially when automated models are expanded in such way (Chaturvedi et al., 2019). In line with our attempt to limit arbitrary choices on feature selection, adding between-channel-connectivity would expand the model with the factorial of 16,674 features and would clearly transcend any current practical computational runtime (García-Martín et al., 2019). Theoretically, our accuracy may yet be further increased by including connectivity- or network features, but the gain in predictive performance is likely limited given the already high accuracy.

Although the ML pipeline treats all patients within one subgroup equally, despite within-group differences in cognitive functioning, the association between the predicted class-probability and actual cognitive performance follows a linear correlation. This trend is predominantly fuelled by the patients on which the model was trained, i.e. high-COG and low-COG patients. Patients classified as int-COG were poorly predicted and no linear trend could be discerned for this subgroup. The final model including all features from the separate cross-validation was inherently not based on 'unseen data' and therefore runs the risk of overfitting, despite several safeguards such as multiple cross-validation runs and Bayesian hyperparameter optimization. This was unavoidable given the small sample size, and the accuracies from the final model are therefore best interpreted as the best approximated maximum, with accuracies from the averaged cross-validation runs as minimum. The risk of overfitting may also partly explain why the model-performance in the int-COG group was ineffective. Other explanations include the limited variability in the int-COG group (by definition, all patients had cognitive scores within 1.5 SD) and variation in cognitive performance within this limited range is likely to occur regardless of the degree of cortical PD pathology and reflect normal variation also found in the otherwise 'normal' population. We emphasize that patients with an 'intermediate' cognition were never included during the initial-model building and therefore constitute a separate class which is rightly unrecognized by the model. Although potentially an interesting group to have further biomarkers on, their cognition is likely more influenced by external factors such as education, motivation, and random effects and less well characterized by pathophysiological changes than the extremer tertiles. The initial ML algorithm is therefore unable to put these patients in a class into which they, clinically speaking, do not belong. Incorporating the intermediate class into the model (as shown in the additional analyses to classify low-COG vs. int-COG and int-COG vs. high-COG) results in slightly lower accuracies, but highly unbalanced sensitivity vs. specificity due to the large class imbalance.

Other than the occipital peak frequency, none of the features retained in the ML pipeline were significantly different between the cognitive classes. This emphasizes the role of a cortical profile of EEG alterations in cognitive functioning and a need to combine multiple EEG-features rather than focussing on a single EEG-biomarker. It is particularly noteworthy that of the 13 retained features, 10 were derived from a Fast Fourier Transformation (FFT) which is typically associated with spectral metrics. Given that the FFT metrics constitute only a fraction of the available metrics from the *tsfresh* feature library, we hypothesize that the neurophysiological profile underlying cognitive alterations is predominantly spectrally-based, and less related to measures of intra-channel connectivity such as entropy or autocorrelation. This is in line with previous literature that showed the importance of spectral metrics compared to connectivity variables, although these metrics were mostly related to inter-channel connectivity (Geraedts et al., 2018a). Within the EEG spectrum, multiple aspects of the FFT appear important, including the real and imaginary parts of the FFT, the skewness and the angle. In terms of localization, the ML pipeline selected features within the parieto-occipital regions (10/13 features), consistent with previous literature on cognition (Babiloni et al., 2015; Geraedts et al., 2018a).

In contrast to previous studies that explored a wide range of cognitive functioning in PD patients, our results focus on PD patients undergoing the screening procedure for DBS. DBS candidates often have a relatively longer disease duration to allow for several treatment options before considering DBS surgery and often have more severe PD symptoms than newly-diagnosed PD patients. Furthermore, severe cognitive impairment is a contraindication for DBS (Geraedts et al., 2019; Lang et al., 2006) and patients with obvious cognitive deficits will not be referred for screening, indicating that the range of cognitive function is likely much smaller in the DBS population than in the global PD population, emphasizing the sensitivity of this ML pipeline.

As with all supervised learning models, the crucial determinant of the models' validity is the correct labelling (either high-COG or low-COG, or another arbitrarily defined label). In this study, an extensive neuropsychological test battery was used to determine cognitive functioning of six consensus-based domains (American Psychiatric Association, 2013), and a derived composite score reflecting global cognition. However, cognitive (dys)function is not a purely binary classification: performance is rated in a spectrum of possible scores and a derived binary classification may be subject to discussion. In this study, classes of cognitive functioning were determined in a data-driven fashion by taking the twenty best- and worst performing patients from the entire cohort. This was an a priori defined classification, as it was deemed unlikely that there would be sufficient DBS candidates with either MCI or PDD. However, it should be noted that both a classification based on the neuropsychological test battery, and cognitive-screening-tests reported previously (Koch et al., 2019), yielded similar model performances suggesting high accuracy regardless of the exact tests used for cognitive profiling.

Our results therefore indicate the utility of using qEEG as complementary biomarker to assess cognitive function, but do not provide an answer towards the pathophysiological mechanism underlying cognitive deficits. We speculate that higher-density source-space setups may provide a better indication of such an underlying mechanism, possibly using Magnetoencephalography (MEG) to better reflect subcortical structures (Bonanni, 2019). However, such an approach would have lower practical utility as it would be more difficult to implement high-density EEG or MEG in routine clinical practice. Nevertheless, this study demonstrates the cortical spatial expansion of the mechanism underlying cognitive impairment.

The ultimate ground truth in terms of clinical impact would be a classification based on long-term postoperative cognitive functioning. This data is however not available, whereas patients with poor preoperative functioning, as identified by the neuropsychological test battery, may be rejected for DBS surgery after screening and thus not contribute to follow-up data. In the high-COG group, 18/20 patients ultimately received DBS (one rejection due to atrophy on the MRI, and one patient opted for gamma-knife surgery instead). In the low-COG group, 12/20 patients received DBS (all rejections due to cognitive impairment). We emphasize that there are other reasons to perform-, or refrain from, DBS apart from cognitive functioning (Geraedts et al., 2019). In terms of predicting future cognitive decline, several previous studies limited to spectral metrics have reported on the utility of EEG (Geraedts et al., 2018a). A low occipital peak frequency (<8.5 Hz) in particular has been associated with a 13-fold higher hazard of developing PDD (Klassen et al., 2011). Given the superiority of our approach in terms of reflecting current cognitive functioning, we hypothesize that our ML pipeline may have similar potential in predicting future cognitive decline as well.

Strengths of our study include the automated ML pipeline which circumvents making arbitrary choices on pre-processing and feature selection, the large number of extracted features, and extensive cognitive profiling on which the initial classification was based. The use of cross-validation warrants the internal validity of our model. To our knowledge, ours is the only cohort of consecutively included DBS candidates with PD with EEG data available in the literature. Given the uniqueness of our cohort, no external validity can therefore be assessed. Despite multiple cross-validation runs, the algorithm was trained on only 20 vs. 20 patients. This constitutes a small sample size to base definitive conclusions on and requires validation in a larger cohort. Nevertheless, our results clearly demonstrate the utility of qEEG during the DBS screening for automated cognitive profiling and the superiority of a compound of EEG features over a single spectral feature.

The classification was based on the most extreme patients with composite scores of six Z-transformed domains. The domains 'Neuropsychiatric functioning' and 'Psychomotoric speed' were paradoxically worse in patients classified as high-COG than low-COG. A possible explanation for this is the younger age in high-COG patients in which PD places a higher burden on daily functioning, despite lower severity of symptoms. However, these factors do not constitute a contra-indication for surgery.

Future studies may confirm the external validity of our model within the population of DBS candidates and evaluate the use of such a ML pipeline on other neurodegenerative diseases with cognitive impairment such as Alzheimer's Disease of Dementia with Lewy Bodies (Dauwan et al., 2016). In such a way, it could be determined whether biomarkers differentiating cognitive subtypes are disease-specific (i.e. different biomarkers for different diseases), or whether there is a neurophysiological compound underlying cognitive impairment across neurodegenerative diseases. As the *ts-fresh* library used by us for feature-extraction is by no means exhaustive, future studies may evaluate the predictive potential of our automated feature-extraction methods on other libraries such as Deep Canonical Correlation Analysis (Andrew et al., 2013) or tslearn (Tavenard et al., 2020), to investigate whether these methods produce similar accuracies and if so, which features are retained. Furthermore, the ultimate goal of the ML pipeline would be to determine its utility as a predictor of cognitive deterioration rather than cross-sectional classification of cognitive functioning.

Strikingly, the model proposed here was originally developed for the automotive industry and applied here to a vastly different research field. This suggests that the origin of the time series, i.e. whether a signal originates from an EEG or from a vehicle, is not important during analyses. We speculate that multidisciplinary approaches such as these may advance healthcare-research and valorise these higher-order analysis-techniques through applications in fundamentally different fields.

We emphasize that currently, the EEG analyses described here are not intended to replace the neuropsychological assessments during the DBS screening and should be seen as complementary. However, these results provide strong evidence of the utility of qEEG as a biomarker for cognitive performance during the DBS screening and may have potential both in clinical practice and for future clinical trials studying disease modifying therapy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the members of the DBS team of LUMC/Haga Teaching Hospital (G.E.L. Hendriks, A. Mosch, R. Zutt, C.F. Hoffmann, N.A. van der Gaag) for patient care and the EEG technicians of the LUMC for their help with the data acquisition.

Funding sources

This work was supported by a grant from the 'Stichting ParkinsonFonds' and the 'Stichting Alkemade-Keuls'.

Author contributions

1. V.J. Geraedts MD MSc: conceived the study and responsible for scientific integrity. Data collection, data analysis, writing the manuscript.
2. M. Koch MSc: data analysis, critical revision of the manuscript.
3. M.F. Contarino MD PhD: conceived the study, critical revision of the manuscript.
4. H.A.M. Middelkoop PhD: data collection, critical revision of the manuscript.
5. H. Wang PhD: data analysis, critical revision of the manuscript.
6. J.J. van Hilten MD PhD: critical revision of the manuscript.
7. T.H.W. Bäck PhD: critical revision of the manuscript.
8. M.R. Tannemaat MD PhD: conceived the study and responsible for its scientific integrity. Critical revision of the manuscript.

Disclosures

V.J. Geraedts reports no disclosures. M. Koch reports no disclosures. M.F. Contarino received grant support from the Stichting ParkinsonFonds, travel support from Boston Scientific, Consultancy-fees (to institution) from Medtronic, has been a member of the scientific advisory board of Medtronic and received unrestricted educational grant support from Medtronic. H.A.M. Middelkoop has nothing to disclose. H. Wang reports no disclosures. J.J. van Hilten received grant support from the Stichting Alkemade Keuls, Stichting ParkinsonFonds, and the Netherlands Organisation for Health Research and Development (ZonMw). T. H.W. Bäck received grant support from the Netherlands Organisation for Scientific Research (NWO). M.R. Tannemaat received grant support from the Stichting Alkemade Keuls.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2021.01.021.

References

Ahlskog JE, Muenter MD. Frequency of levodopa-related dyskinesias and motor fluctuations as estimated from the cumulative literature. Mov Disord 2001;16(3):448–58.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 5th ed. Arlington: VA: American Psychiatric Publishing; 2013.

Andrew G, Arora R, Bilmes J, Livescu K. Deep Canonical Correlation Analysis. In: Sanjoy D, David M, editors. Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR. p. 1247–55.

Babiloni C, Del Percio C, Boccardi M, Lizio R, Lopez S, Carducci F, et al. Occipital sources of resting-state alpha rhythms are related to local gray matter density in subjects with amnesic mild cognitive impairment and Alzheimer's disease. Neurobiol Aging 2015;36(2):556–70.

Beck AT, Steer RA, Ball R, Ranieri W. Comparison of beck depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996;67(3):588–97.

Betrouni N, Delval A, Chaton L, Defebvre L, Duits A, Moonen A, et al. Electroencephalography-based machine learning for cognitive profiling in Parkinson's disease: preliminary results. Mov Disord 2019;34(2):210–7.

Bonanni L. The democratic aspect of machine learning: limitations and opportunities for Parkinson's disease. Mov Disord 2019;34(2):164–6.

Burle B, Spieser L, Roger C, Casini L, Hasbroucq T, Vidal F. Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. Int J Psychophysiol 2015;97(3):210–20.

Chaturvedi M, Bogaarts JG, Kozak Cozac VV, Hatz F, Gschwandtner U, Meyer A, et al. Phase lag index and spectral power as QEEG features for identification of patients with mild cognitive impairment in Parkinson's disease. Clin Neurophysiol 2019;130(10):1937–44.

Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh – A Python package). Neurocomputing 2018;307:72–7.

Christ M, Kempa-Liehr AW, Feindt M. Distributed and parallel time series feature extraction for industrial big data applications. arXiv e-prints 2016.

Contarino MF, Daniele A, Sibilia AH, Romito LM, Bentivoglio AR, Gainotti G, et al. Cognitive outcome 5 years after bilateral chronic stimulation of subthalamic nucleus in patients with Parkinson's disease. J Neurol Neurosurg Psychiatry 2007;78(3):248–52.

Dauwan M, van der Zande JJ, van Dellen E, Sommer IEC, Scheltens P, Lemstra AW, et al. Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease. Alzheimers Dement (Amst) 2016;4:99–106.

Dekker R, Mulder JL, Dekker PH. De ontwikkeling van vijf nieuwe Nederlandstalige tests. Leiden: PITS; 2007.

Deuschl G, Agid Y. Subthalamic neurostimulation for Parkinson's disease with early fluctuations: balancing the risks and benefits. Lancet Neurol 2013;12(10):1025–34.

Drapier D, Drapier S, Sauleau P, Haegelen C, Raoul S, Biseul I, et al. Does subthalamic nucleus stimulation induce apathy in Parkinson's disease? J Neurol 2006;253(8):1083–91.

Duckworth AL, Quinn PD, Lynam DR, Loeber R, Stouthamer-Loeber M. Role of test motivation in intelligence testing. Proc Natl Acad Sci USA 2011;108(19):7716–20.

Duckworth AL, Yeager DS. Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. Educ Res 2015;44(4):237–51.

Duncan JS. Conventional and clinimetric approahces to individualization of antiepileptic drug therapy. In: Meinardi H, Cramer JA, Baker GA, da Silva AM (editors). Quantitative assessment in epilepsy care. Porto, Portugal: Springer Science+Business Media, LLC; 1993.

García-Martín E, Rodrigues CF, Riley G, Grahn H. Estimation of energy consumption in machine learning. J Parallel Distrib Comput 2019;134:75–88.

Geraedts VJ, Boon LI, Marinus J, Gouw AA, van Hilten JJ, Stam CJ, et al. Clinical correlates of quantitative EEG in Parkinson disease: a systematic review. Neurology 2018a;91(19):871–83.

Geraedts VJ, Kuijf ML, van Hilten JJ, Marinus J, Oosterloo M, Contarino MF. Selecting candidates for Deep Brain Stimulation in Parkinson's disease: the role of patients' expectations. Parkinsonism Relat Disord 2019;66:207–11.

Geraedts VJ, Marinus J, Gouw AA, Mosch A, Stam CJ, van Hilten JJ, et al. Quantitative EEG reflects non-dopaminergic disease severity in Parkinson's disease. Clin Neurophysiol 2018b;129(8):1748–55.

Geron A. Hands-on Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to build Intelligent Systems. Sebastopol, CA: O'Reilly Media; 2017.

Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, 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 2008;23(15):2129–70.

Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, New York; 2009.

Hjorth B. Source derivation simplifies topographical EEG interpretation. Am J EEG Technol 1980;20(3):121–32.

Huppert FA, Brayne C, Gill C, Paykel ES, Beardsall L. CAMCOG–a concise neuropsychological test to assist dementia diagnosis: socio-demographic determinants in an elderly population sample. Br J Clin Psychol 1995;34:529–41.

Klassen BT, Hentz JG, Shill HA, Driver-Dunckley E, Evidente VG, Sabbagh MN, et al. Quantitative EEG as a predictive biomarker for Parkinson disease dementia. Neurology 2011;77(2):118–24.

Koch M, Bäck T. Machine Learning for Predicting the Impact Point of a Low Speed Vehicle Crash. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). p. 1432–7.

Koch M, Geraedts V, Wang H, Tannemaat MR, Bäck T. Automated Machine Learning for EEG-Based Classification of Parkinson's Disease Patients. In: 2019 IEEE International Conference on Big Data. Los Angeles; 2019. p. 4845-52.

Koch M, Wang H, Bäck T. Machine Learning for Predicting the Damaged Parts of a Low Speed Vehicle Crash. In: 13th International Conference on Digital Information Management; 2018. p. 179-84.

Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw 2010;36(11):1–13.

Lang AE, Houeto JL, Krack P, Kubu C, Lyons KE, Moro E, et al. Deep brain stimulation: preoperative issues. Mov Disord 2006;21(Suppl 14):S171–96.

Litvan I, Goldman JG, Troster AI, Schmand BA, Weintraub D, Petersen RC, et al. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Mov Disord 2012;27(3):349–56.

Okun MS, Gallo BV, Mandybur G, Jagid J, Foote KD, Revilla FJ, et al. Subthalamic deep brain stimulation with a constant-current device in Parkinson's disease: an open-label randomised controlled trial. Lancet Neurol 2012;11(2):140–9.

Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015;30(12):1591–601.

Richardson JTE. Measures of short-term memory: a historical review. Cortex 2007;43(5):635–50.

Scarpina F, Tagini S. The Stroop Color and Word Test. Front Psychol 2017;8:557-.

Smeding HM, Speelman JD, Huizenga HM, Schuurman PR, Schmand B. Predictors of cognitive and psychosocial outcome after STN DBS in Parkinson's Disease. J Neurol Neurosurg Psychiatry 2011;82(7):754–60.

Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol 2018;103:131–3.

Tavenard R, Faouzi J, Vandewiele G, Divo F, Androz G, Holtz C, et al. Tslearn, A machine learning toolkit for time series data. J Mach Learn Res 2020;21(118):1–6.

Tombaugh TN. Trail Making Test A and B: normative data stratified by age and education. Arch Clin Neuropsychol 2004;19(2):203–14.

Utianski RL, Caviness JN, van Straaten ECW, Beach TG, Dugger BN, Shill HA, et al. Graph theory network function in parkinson's disease assessed with electroencephalography. Clin Neurophysiol 2016;127(5):2228–36.

Vakil E, Blachstein H. Rey Auditory-Verbal Learning Test: structure analysis. J Clin Psychol 1993;49(6):883–90.

van der Heeden JF, Marinus J, Martinez-Martin P, van Hilten JJ. Evaluation of severity of predominantly non-dopaminergic symptoms in Parkinson's disease: the SENS-PD scale. Parkinsonism Relat Disord 2016;25:39–44.

Wang H, Emmerich M, Bäck T. Cooling Strategies for the Moment-Generating Function in Bayesian Global Optimization. 2018 IEEE Congress on Evolutionary Computation (CEC); 2018. p. 1-8.

Wang H, Stein Bv, Emmerich M, Back T. A new acquisition function for Bayesian optimization based on the moment-generating function. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017. p. 507-12.

Weaver FM, Follett K, Stern M, Hur K, Harris C, Marks WJ, et al. Bilateral deep brain stimulation vs best medical therapy for patients with advanced Parkinson disease a randomized controlled trial. JAMA 2009;301(1):63–73.

Wechsler D. Wechsler memory scale. San Antonio, TX, US: Psychological Corporation; 1945.

Yang K, van der Blom K, Bäck T, Emmerich M. Towards single- and multiobjective Bayesian global optimization for mixed integer problems. AIP Conf Proc 2019;2070(1):020044.

Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361–70.