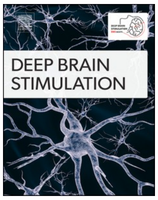
Contents lists available at ScienceDirect

Deep Brain Stimulation

journal homepage: www.journals.elsevier.com/deep-brain-stimulation



Prediction of STN-DBS outcome in Parkinson's disease using machine learning

Laurens A. Biesheuvel [a,b,1,*], Jesús Fuentes [c,1], Rob M.A. de Bie [a], Bernadette C.M. van Wijk [a,d], P. Rick Schuurman [e], Andreas Husch [c], Jorge Goncalves [c,f], Martijn Beudel [a]

[a] *Department of Neurology, Amsterdam Neuroscience Institute, Amsterdam University Medical Center, , Amsterdam, the Netherlands*
[b] *Department of Intensive Care Medicine, Amsterdam University Medical Center, , Amsterdam, the Netherlands*
[c] *Luxembourg Centre for Systems Biomedicine, University of Luxembourg, , Belvaux, L-4367, Luxembourg*
[d] *Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, , Amsterdam 1081 BT, the Netherlands*
[e] *Department of Neurosurgery, Amsterdam Neuroscience Institute, Amsterdam University Medical Center, Amsterdam, the Netherlands*
[f] *Department of Plant Sciences, Cambridge University, , Cambridge CB2 3EA, United Kingdom*

ARTICLE INFO

Keywords:
DBS
Machine learning
Ai
Statistics

ABSTRACT

Deep brain stimulation (DBS) targeting the subthalamic nucleus (STN) is an established therapy for advanced Parkinson's disease (PD), but outcomes vary significantly among patients. Using a dataset of 408–420 PD patients (depending on outcome), we developed machine learning models to predict outcomes of STN-DBS based on preoperative clinical markers. Regression models predicted scores on the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) part III, and subscores for Tremor, Axial symptoms, and Bradykinesia & Rigidity. The models achieved root mean square errors (RMSE) of 9.1, 2.6, 2.5, and 5.3, respectively. These results demonstrate the models' ability to provide accurate predictions despite the heterogeneity of PD. This approach refines patient selection by forecasting postoperative outcomes and enables personalized treatment planning. Future iterations will explore additional predictors, such as neuroimaging data, to further improve model performance and support clinical decision-making in DBS therapy. This study advances the use of machine learning in predictive medicine for PD.

Introduction

Deep brain stimulation (DBS) of the subthalamic nucleus (STN) is a prominent surgical intervention for managing the symptoms of advanced Parkinson's disease (PD). The efficacy of this therapeutic approach is well-documented, with optimized treatment yielding notable clinical outcomes [1,2]. However, the clinical improvement varies widely among patients [3–5], with some experiencing minimal or no benefit despite careful patient selection [6]. Additionally, there are risks of surgical complications such as intracranial hemorrhage and stimulation-induced side effects such as dysarthria [7,8]. Hence, identifying patients who are more likely to have a favorable outcome after DBS is crucial.

Key predictors for a good outcome in PD are younger age and a higher preoperative levodopa responsiveness [9–18]. However, the consistency of these relationships across studies varies, and identifying the most informative predictors remains challenging due to incomplete datasets and statistical dependencies among variables [15,19–23]. Furthermore, there is evidence that preoperative depressive symptoms, dispositional optimism, and psychosocial factors significantly influence motor and quality of life improvements post-DBS [24–27]. This warrants the inclusion of non-motor parameters such as apathy (Starkstein Apathy Scores, SAS), quality of life (Parkinson's Disease Questionnaire-39, PDQ39), and activities of daily living (ADL). Nevertheless, factors such as precise lead positioning and postoperative programming are also crucial and will be considered in future studies.

Machine learning (ML) is a promising approach for predicting patient responsiveness to STN-DBS. ML methods enable algorithms to identify patterns from data to learn to predict outcomes with minimal human intervention. ML can address the complexities inherent in STN-

[*] Corresponding author at: Department of Neurology, Amsterdam Neuroscience Institute, Amsterdam University Medical Center, Amsterdam, the Netherlands.
E-mail address: l.biesheuvel@amsterdamumc.nl (L.A. Biesheuvel).
[1] Contributed equally.

https://doi.org/10.1016/j.jdbs.2025.10.002
Received 19 July 2025; Received in revised form 12 October 2025; Accepted 16 October 2025
Available online 16 October 2025

DBS outcome prediction, like the wide variability in disease manifestations, disease progression rates, individual responses to treatment and the noisy nature of clinical data. Supervised learning algorithms, which learn from data that includes both input features and labeled output, are particularly valuable to predict outcomes for new, unseen data [28]. Nevertheless, many current medical models have inconsistent accuracy when predicting outcomes for individual patients. This can be attributed to their inability to effectively navigate the high dimensionality and heterogeneity inherent in clinical and demographic data [28,29]. In line with this, Cerasa emphasized that translating ML into PD clinical practice requires methodological rigor and high-quality labels, that ML should support rather than replace clinician judgment, and that claims should be tempered until external validation and reproducibility barriers are addressed [32].

Previous studies have used ML to predict DBS outcome for PD with differing endpoints and sample sizes. Habets et al. [6] constructed an ML model using retrospective data from 89 patients. While their study employed a multi-faceted outcome variable and showed high performance in terms of evaluation metrics, their approach yielded a binary classification model that relied on predefined thresholds to differentiate between favorable responders and weak responders with an accuracy of 78 %. While the original study used a modest single-centre dataset, subsequent external, multicentre validation has supported the model's generalisability and clinical relevance [30]. In a different effort to predict DBS outcomes, Liu et al. [31] analyzed a cohort of 33 PD patients undergoing bilateral STN-DBS. Their study addressed the heterogeneity of iron distribution within the substantia nigra and its potential predictive value for motor outcomes post-DBS. Employing an ML radiomics-based approach to analyze preoperative quantitative susceptibility mapping as a binary classification model, the authors developed predictive models that were able to distinguish favorable responders and weak responders with accuracies up to 82 % for global motor and rigidity outcomes.

Our study expands on these foundational works by addressing identified gaps and recognising the strengths of both approaches. Similar to Habets et al. [6], we focussed on preoperative clinical predictors but utilized a much larger database and a refined modeling process. Contrary to the binary categorization followed by both studies, our methodology adopts a regression-based analysis to predict continuous preoperative variables and postoperative clinical outcomes, aiming for a more fine-grained prediction of patient outcomes following STN-DBS. This strategy avoids the biases in classifying patients based on arbitrary cutoffs and aligns with the objective of predicting actual post-surgical improvements. Hence, pure postoperative information is used to define the output variables, which prevents preoperative data from contaminating predictions. Therefore, our models specifically predict absolute raw Movement Disorders Society Unified Parkinson's Disease Rating Scale motor (MDS-UPDRS-III) scores, not relative treatment improvements, enabling a more informed dialogue between healthcare professionals and patients.

We developed and applied four regression models to predict key postoperative clinical indicators in PD management. These include the total MDS-UPDRS-III score evaluated when patients are Off medication (Off Med) and on Deep Brain Stimulation (On DBS), alongside total Tremor, Total Axial symptoms, and Bradykinesia & Rigidity MDS-UPDRS-III subscores under the same conditions.

Methods

Since 2018, demographic and clinical data from patients undergoing STN-DBS for PD are being collected at the Amsterdam University Medical Center under local regulations with a waiver for active consent. Of 1122 records, 17 (1.5 %) withheld consent, leaving 1105 eligible. For each target outcome we required the essential inputs—baseline MDS-UPDRS-III Off and On (or the relevant subscores) and the corresponding postoperative Off-medication/On-DBS score; records missing any of these were excluded. Exclusions and final analytic sizes were: Total MDS-UPDRS-III: 697 excluded (final n = 408); Tremor: 685 (n = 420); Axial: 686 (n = 419); Bradykinesia & Rigidity: 689 (n = 416). The most frequent missing item was the postoperative Off-medication/On-DBS assessment (absent in 670–672 records, depending on outcome). Collected data include demographics, disease characteristics, medication history, and PD symptom severity. We conducted a retrospective cohort study to explore relevant predictors and assess the potential of ML models in predicting patient outcomes after STN-DBS. The typical postoperative evaluation occurred 12 months after STN-DBS, although some variability existed. We modeled timing heterogeneity by including 'Days between screening and follow-up' as a candidate predictor in all models.

We selected preoperative predictors that represent the patient's condition and treatment background and categorized them into several groups. Demographic and disease-specific characteristics included: Age, Sex, Disease Duration, and the presence of Impulse Control Disorder. Treatment and medication-related factors comprised: Days between screening and postoperative follow-up, and Total Levodopa Equivalent Dose. We further incorporated detailed clinical assessments obtained during screening, focusing on various MDS-UPDRS scores. These encompassed: Total MDS-UPDRS-I, MDS-UPDRS-II On Med, MDS-UPDRS-III Off Med, MDS-UPDRS-III On Med, along with subscores for Bradykinesia & Rigidity (both On Med and Off Med), Tremor (both On Med and Off Med), and Axial symptoms (both On Med and Off Med) symptoms, and the Total MDS-UPDRS-IV score. Additionally, the improvement percentages in MDS-UPDRS-III post-levodopa administration were evaluated, including overall MDS-UPDRS-III improvement and specific improvements in Bradykinesia & Rigidity, Tremor, and Axial symptoms. Lastly, the patients' quality of life and overall well-being were measured using the PDQ39 score, while apathy was determined using the SAS scores. Other variables included the percentage of the waking day with Dyskinesias, Off time, and Off time with Dystonia. The targeted predictive outcomes were postoperative Off Med On DBS Total MDS-UPDRS-III, Tremor subscore (items 3.15–3.17), Axial symptom subscore (items 3.9–3.13), and Bradykinesia and Rigidity subscore (items 3.3–3.8). We chose this set because it is routinely collected at the DBS screening visit, which makes the models portable across centres. We chose the Off Med On DBS state to isolate the stimulation effect and reduce confounding from short-term dopaminergic fluctuations and postoperative medication adjustments, providing a standardized postoperative motor assessment.

Data analysis was performed in Python 3.9 using Scikit-Learn 1.3 [33], SHAP 0.44 [34], Pandas 2.1 [35], Numpy 1.26 [36], Matplotlib 3.8 [37], Optuna 3.4 [38], and Imbalanced-Learn 0.11 [39]. We randomly allocated patients into a training set (80 %) and a testing set (20 %) for evaluation.

The full data processing, training, optimization and evaluation strategy we employed is depicted in Figs. 1, 2. The approach started with feature selection to narrow down the predictors to the most relevant, mitigating overfitting and promoting generalizable pattern learning. This was performed using Recursive Feature Elimination with Cross-Validation (RFECV) using a pipeline that employed a linear regression model. RFECV was embedded in a 5-fold cross-validation to mitigate selection variance. Missing data in the training set were imputed using Scikit-learn's IterativeImputer, assuming missing at random conditional on observed covariates, with imputation performed within CV folds to avoid leakage. Features were scaled using Robust Scaler to ensure equal processing and faster training. Outliers were addressed with the Elliptic Envelope method [40], implemented as the EllipticEnvelope class in [33]. The elliptic envelope was configured with a contamination parameter of 0.10, indicating that we anticipate approximately 10 % of the data points to be outliers. A contamination fraction of 0.10 was chosen to deal with measurement errors in big clinical datasets while avoiding trimming genuine clinical heterogeneity.

Three candidate algorithms were selected as our primary tools for

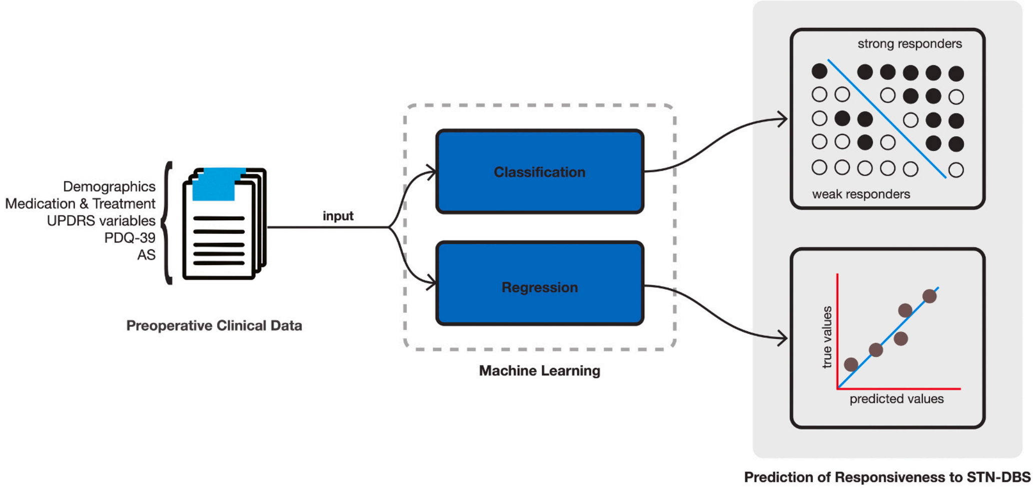


Fig. 1. Overview of classification and regression methods to predict possible candidates to STN-DBS. *Abbreviations: STN-DBS, subthalamic nucleus deep brain stimulation; MDS-UPDRS, Movement Disorder Society–Unified Parkinson's Disease Rating Scale; PDQ-39, Parkinson's Disease Questionnaire-39; AS, Apathy Scale.*

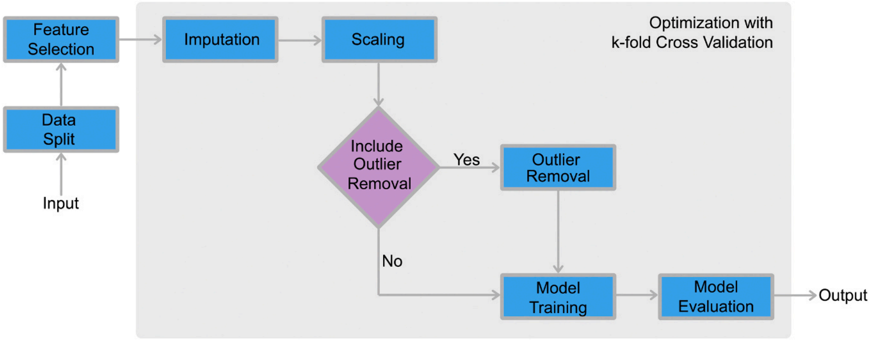


Fig. 2. Schematic representation of the ML pipeline used in the study. This diagram illustrates the sequential process from 'Input' to 'Output', encompassing key stages in the algorithm.

regression analysis: Random Forest, Linear Regression, and LASSO, each chosen for its distinctive advantages and potential limitations within the regression framework. The Random Forest regressor was selected for its robustness and capacity to capture complex, nonlinear relationships. Linear Regression, in contrast, offers a more direct approach, predicated on assumptions that are best suited for analyzing simpler, linear relationships. LASSO, with its regularization technique and built-in feature selection capabilities, could aid in the reduction of model complexity and mitigation of overfitting, thereby enhancing model performance. The adoption of a diverse array of candidate algorithms is a pragmatic approach to promote the determination of the most suitable model for STN-DBS outcome prediction. As a baseline check, we also compared simple correlation analyses, as shown in Tables 1 and 2.

We prioritised ordinary linear regression, LASSO, and Random Forests to balance interpretability and nonlinearity within the constraints of a few hundred training cases. Linear models provide transparent coefficient effects and calibrated predictions; LASSO offers embedded regularisation to manage overfitting in the presence of correlated clinical measures; Random Forests flexibly capture interactions and nonlinearities with robust defaults. More complex ensembles and neural networks were not pursued here to avoid excessive parameterisation and to maintain accessible clinical interpretability; exploratory comparisons are reserved for future external validation work.

In order to optimize the performance of each model, we utilized Optuna, a framework for automating the process of hyperparameter search, a process to determine the best settings that influence the models' learning behavior. The assessed hyperparameters included the number of trees and maximum depth for the Random Forest models, and the alpha parameter for the LASSO models. In addition, the added value of the outlier removal process was evaluated by enabling or disabling it during the fine-tuning process, and its inclusion/exclusion was part of model selection based on cross-validated score. The impact of these hyperparameters on model performance was determined using 5-fold cross-validation on the training set, with the goal of minimizing the mean squared error (MSE). A total of 100 trials were performed using Optuna's Tree-structured Parzen Estimator (TPESampler).

Each of the four regression models—targeting different clinical indicators of PD—was trained and optimized independently. This approach ensured that the specificities of each target variable were adequately addressed. Upon determination of the best model configurations, we trained a final regression model for each outcome using the full training dataset, as opposed to the cross validation subsets used for hyperparameter tuning. Subsequently, we evaluated the final performance of our models on the separate test dataset by calculating the root mean square error (RMSE).

To assess how the model could be improved in future iterations, we generated learning curves for our final ML models. Learning curves plot the model's performance against the training set size to visualize the improvement in prediction quality as more training data is incorporated. The main purpose of analyzing learning curves is to identify whether the model is suffering from high bias or high variance. High bias indicates that the model potentially oversimplifies patterns, while high variance suggests an overfitted model that captures noisy data as relevant signals. This analysis could help guide further adjustments to the model, such as

Table 1

Clinical Characteristics and Correlations with MDS-UPDRS-III Subscores. This table presents both categorical and continuous clinical features of the study population, along with their associations with the MDS-UPDRS-III motor examination subscores, which assess total motor impairment, tremor, axial symptoms, and rigidity/bradykinesia. For categorical features (e.g., sex, impulse control disorder), the number of patients (n) and percentages are reported, along with the Mean ± SD of each MDS-UPDRS-III subscore. For continuous variables (e.g., age, disease duration), the overall Mean ± SD of each feature is provided, along with the Pearson correlation coefficients between the feature and the MDS-UPDRS-III subscores. Counts and percentages are based on available data per variable; denominators vary across rows due to missingness.

Feature	Category	n (%)	MDS-UPDRS-III Total Mean ± SD	MDS-UPDRS-III Tremor Mean ± SD	MDS-UPDRS-III Axial Mean ± SD	MDS-UPDRS-III Rigidity + Bradykinesia Mean ± SD
Sex	Male	272 (64.3 %)	23.4 ± 14.4	3.3 ± 4.3	2.6 ± 3.5	13.4 ± 8.9
	Female	151 (35.7 %)	20.4 ± 13.2	2.7 ± 3.8	2.2 ± 3.2	11.9 ± 8.8
Impulse control disorder	No	245 (96.8 %)	21.4 ± 14.6	2.7 ± 4.0	2.4 ± 3.6	12.6 ± 9.3
	Yes	8 (3.2 %)	22.3 ± 11.3	5.3 ± 4.2	2.6 ± 3.2	9.6 ± 10.1
	Mean ± SD		Correlation with MDS-UPDRS-III Total	Correlation with MDS-UPDRS-III Tremor	Correlation with MDS-UPDRS-III Axial	Correlation with MDS-UPDRS-III Rigidity + Bradykinesia
Age	62.7 ± 8.4		0.21	0.11	0.29	0.09
Disease duration	10.3 ± 5.0		0.10	0.00	0.14	0.09
Days between screening and follow up	361.7 ± 143.2		0.01	−0.05	0.08	−0.03
Total Levodopa Equivalent Dose	1390.5 ± 652.5		0.03	−0.04	0.15	0.00
Total MDS-UPDRS-I score	12.8 ± 5.6		0.14	−0.05	0.17	0.15
Total MDS-UPDRS-II On score	16.8 ± 6.9		0.30	0.00	0.26	0.27
Total MDS-UPDRS-III Off score	48.0 ± 14.6		0.45	0.24	0.33	0.33
Total bradykinesia + rigidity Off score (MDS-UPDRS-III subscore)	27.9 ± 8.8		0.41	0.01	0.26	0.41
Total tremor Off score (MDS-UPDRS-III subscore)	4.6 ± 5.1		0.11	0.58	−0.06	−0.07
Axial Off score (MDS-UPDRS-III subscore)	5.8 ± 4.2		0.31	−0.11	0.59	0.23
Total bradykinesia + rigidity On score (MDS-UPDRS-III subscore)	12.9 ± 7.4		0.48	0.05	0.32	0.49
Total MDS-UPDRS-III On score	20.5 ± 10.7		0.52	0.17	0.40	0.46
Total tremor On score (MDS-UPDRS-III subscore)	0.7 ± 1.8		0.14	0.49	−0.01	0.01
Axial On score (MDS-UPDRS-III subscore)	2.3 ± 2.4		0.43	0.02	0.63	0.32
% MDS-UPDRS-III improvement after dopamine	57.4 ± 17.7		−0.31	−0.04	−0.25	−0.31
% Total bradykinesia + rigidity improvement after dopamine (MDS-UPDRS-III subscore)	54.0 ± 20.9		−0.31	−0.08	−0.22	−0.31
% Total tremor improvement after dopamine (MDS-UPDRS-III subscore)	88.2 ± 23.1		−0.14	−0.29	−0.01	−0.07
% Axial improvement after dopamine (MDS-UPDRS-III subscore)	57.8 ± 35.9		−0.19	−0.14	−0.18	−0.14
Total preoperative MDS-UPDRS-IV score	10.6 ± 3.6		0.07	−0.15	0.17	0.11
% of waking day dyskinesias present	28.2 ± 26.2		−0.01	−0.08	−0.02	0.03
% of waking day Off	27.7 ± 17.9		0.08	0.04	0.06	0.08
% of Off time with dystonia	25.5 ± 35.6		0.00	0.00	0.02	0.02
SAS score	6.6 ± 5.0		0.15	0.05	0.06	0.16
PDQ−39 score	31.5 ± 11.6		0.19	−0.06	0.22	0.21

altering its complexity. Moreover, learning curves can help indicate when the model's performance has plateaued, suggesting that adding more samples to the training set may offer limited beneficial impact.

Finally, we employed the SHAP (SHapley Additive exPlanations) framework to examine the individual contributions of features to the model's predictions. SHAP values provide an understanding of the impact each feature has on the output, enabling us to interpret the model's decision-making process more comprehensively.

Data sharing

A synthetic version of the dataset is available upon reasonable request.

The full code of the analysis pipeline is publicly available at: https://github.com/lbiesheuvel/DBSPredict

Results

Of the initial cohort of 1105 patients, the number of patients included in each ML model analysis varied due to the availability of data required for calculating outcome. 408 patients were included for the Total MDS-UPDRS-III Model, 420 for the Tremor Model, 419 for the Axial Model, and 416 for the Bradykinesia + Rigidity Model. In Table 1, we present the results of descriptive analytics applied to all included patients. The mean age of patients in the database was 62.7 ± 8.4 years. The mean duration of PD at the time of screening was 10.3 ± 5.0 years. The study population exhibited a gender imbalance, predominantly comprising males (64.3 %) over females (35.7 %), which reflects the known higher prevalence of PD in males.

After applying the feature selection algorithm, the list of predictors was reduced to those highlighted in Table 2. For the Total MDS-UPDRS-III prediction model, the optimal features were Age, Total MDS-UPDRS-III Off score and Total MDS-UPDRS-III On score. For the Tremor model, the optimal predictors were Total Tremor Off score, Percentage of MDS-

Table 2
This table provides an overview of the performance of each regression model predicting MDS-UPDRS-III subscores (total, tremor, axial, and bradykinesia + rigidity). The number of training and testing samples, model type, RMSE, and R-squared are presented along with the clinical features included in each model. These features, such as the Total MDS-UPDRS-III Off score and Axial On score, were chosen based on their relevance to predicting specific motor outcomes in Parkinson's disease.

Model	Training Samples	Testing Samples	Best Model Type	RMSE (95 % CI)	R^2 (95 % CI)	Included Features
Total MDS-UPDRS-III Model	326	82	Linear Regression	9.1 (7.4–10.8)	0.30 (0.09–0.45)	Age, Total MDS-UPDRS-III Off score and Total MDS-UPDRS-III On score.
Tremor Model	336	84	Random Forest Regressor	2.6 (1.9–3.3)	0.29 (0.12–0.47)	Total Tremor Off score, Percentage of MDS-UPDRS-III Improvement After Dopamine, and Percentage of Total Bradykinesia & Rigidity Improvement After Dopamine.
Axial Model	335	84	Linear Regression	2.5 (2.0–3.0)	0.31 (0.08–0.50)	Age, Total MDS-UPDRS-III Off score, Total Bradykinesia & Rigidity Off score, Total Tremor Off score, Axial On score and Percentage of Axial Improvement After Dopamine.
Bradykinesia + Rigidity Model	332	84	Lasso	5.3 (4.5–6.0)	0.25 (−0.07–0.47)	Age, Total MDS-UPDRS-II On score, Total MDS-UPDRS-III Off score, Total Tremor Off score, Axial Off score, Total Bradykinesia & Rigidity On score, Axial On score, Percentage of Total Bradykinesia & Rigidity Improvement After Dopamine, Percentage of Axial Improvement After Dopamine and SAS score.

Abbreviations: MDS-UPDRS, Movement Disorder Society–Unified Parkinson's Disease Rating Scale; RMSE, root mean square error; R^2, coefficient of determination; LASSO, least absolute shrinkage and selection operator; DBS, deep brain stimulation; PD, Parkinson's disease.

UPDRS-III Improvement After Dopamine, and Percentage of Total Bradykinesia & Rigidity Improvement After Dopamine. For the Axial symptom model, the predictors were Age, Total MDS-UPDRS-III Off score, Total Bradykinesia & Rigidity Off score, Total Tremor Off score, Axial On score and Percentage of Axial Improvement After Dopamine. Finally, the selected features for the Bradykinesia & Rigidity model were Age, Total MDS-UPDRS-II On score, Total MDS-UPDRS-III Off score, Total Tremor Off score, Axial Off score, Total Bradykinesia & Rigidity On score, Axial On score, Percentage of Total Bradykinesia & Rigidity Improvement After Dopamine, Percentage of Axial Improvement After Dopamine and SAS score. 'Days between screening and follow up' was not retained by RFECV for any outcome, indicating limited incremental predictive value. Thus, timing variability likely did not drive between-patient differences captured by the models.

The final model evaluation, following hyperparameter tuning, is detailed in Table 2. Additionally, the prediction evaluations and the learning curves for each model are depicted in the left and right column panels of Fig. 3, respectively. For both the total MDS-UPDRS-III model and the Axial symptom model, linear regression emerged as the best model type. Conversely, the Random Forest Regressor and LASSO were identified as the optimal model types for the Tremor model and the Bradykinesia & Rigidity model, respectively. The RMSE on the testing set was 9.1 (95 % CI: 7.4 – 10.8) for the Total MDS-UPDRS-III model; 2.6 (95 % CI: 1.9 – 3.3) for the Tremor model; 2.5 (95 % CI: 2.0 – 3.0) for the Axial model; and 5.3 (95 % CI: 4.5 – 6.0) for the Bradykinesia & Rigidity model. The corresponding R-squared values were: 0.30 (95 % CI: 0.09 – 0.45) for the Total MDS-UPDRS-III model; 0.29 (95 % CI: 0.12 – 0.47) for the Tremor model; 0.31 (95 % CI: 0.08 – 0.5) for the Axial model; and 0.25 (95 % CI: −0.07 – 0.47) for the Bradykinesia & Rigidity model. The RMSE can be viewed as the average deviation (in MDS-UPDRS-III points) between predicted and actual postoperative scores. Hence, an RMSE of about 9 points for the total MDS-UPDRS III suggests that, on average, a new patient's predicted motor score could be within roughly ±9 points of the actual outcome.

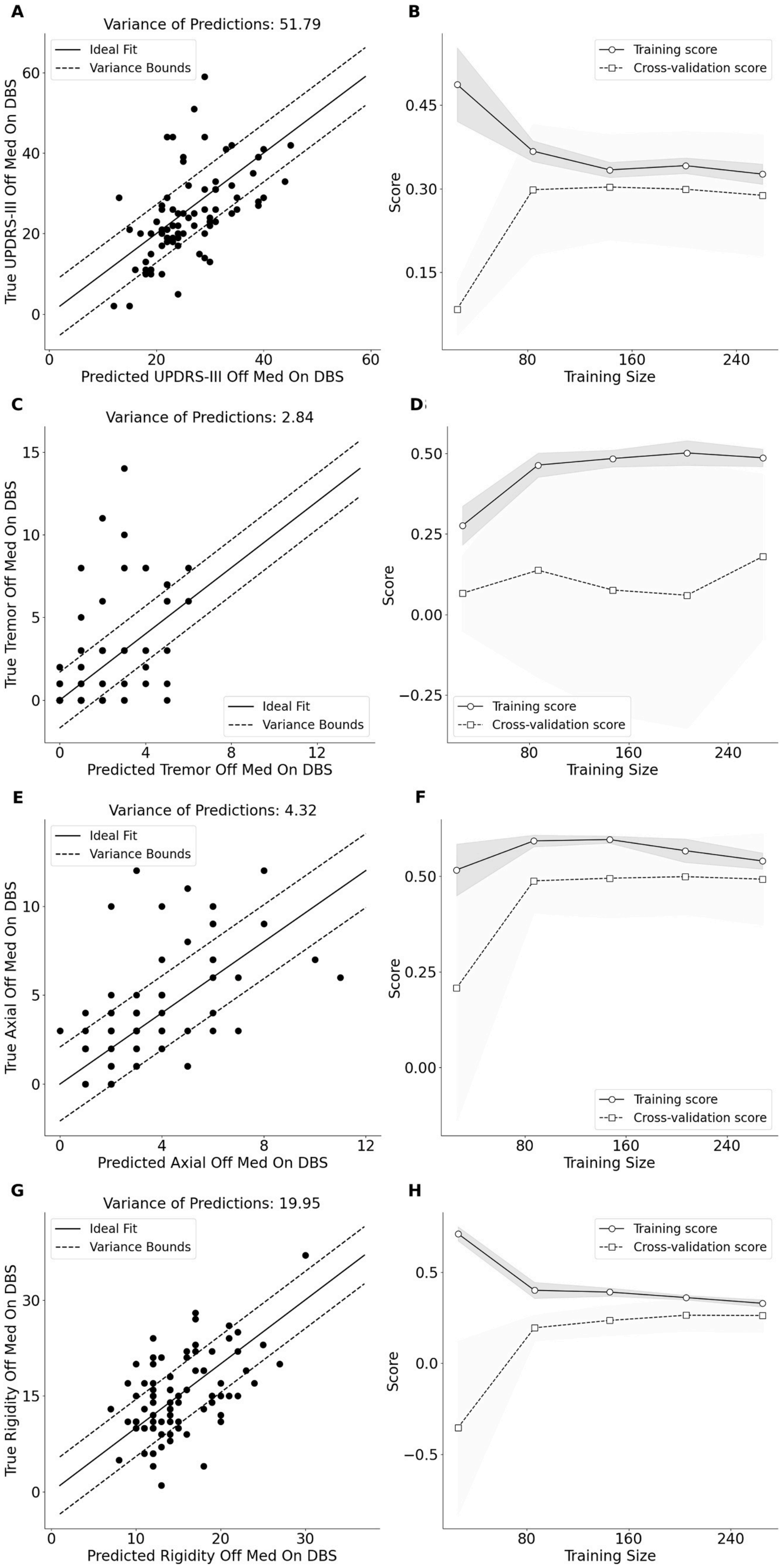
When looking at the medical records of five patients for whom the predictions of total MDS-UPDRS-III improvement were markedly inaccurate, several factors were identified. These included the use of Duodopa therapy, suboptimal contact selection, anxiety during the follow-up scoring, and a (paradoxical) improvement from baseline also in Off-Off and On-On, possibly indicating sub-optimal initial screening conditions. For these cases the discrepancies between the predicted and actual improvements were significant, as reflected in the RMSE values: for actual improvements of 59.0 and 51.0, the predictions were markedly lower at 29.0 and 27.0, respectively; for two instances of actual improvements of 44.0, the predictions underestimated the improvements at 22.0 and 23.0; and most strikingly, for an actual improvement of 5.0, the prediction erroneously suggested a much higher improvement of 24.0.

From examining the learning curves of the final models, as presented in Panels B, D, F, and H of Fig. 3, it becomes evident that the learning progression for the total MDS-UPDRS-III model, the Axial model, and the Bradykinesia & Rigidity model reached a plateau. This observation suggests that these models have attained their optimal learning capacity with the given set of variables, implying that incorporating more training data is unlikely to yield significant performance improvements. By contrast, the learning curve associated with the Tremor model indicates that its performance could be further enhanced with the addition of more training data. These patterns suggest that additional clinical cases alone may offer diminishing returns for Total, Axial, and Bradykinesia & Rigidity outcomes, whereas Tremor appears data-limited; further gains for all models are more likely to come from new predictors that encode anatomy and programming (e.g., lead location, active contact selection, and stimulation parameters).

The summary of the SHAP analysis results is depicted in Figures S1-8. In the model predicting the Total MDS-UPDRS-III outcome, the most influential factors were found to be the preoperative MDS-UPDRS-III scores in both On and Off states, with Age having a comparatively smaller impact, as illustrated in Figure S1-S2. The Tremor model's predictions were primarily influenced by the preoperative Tremor Off scores, with the MDS-UPDRS-III and the Bradykinesia & Rigidity dopamine responsiveness also playing significant roles. Regarding the Axial and Bradykinesia & Rigidity models, a more complex interplay of feature importance was observed. The Axial model showed the highest SHAP values for the Axial On score and the Total MDS-UPDRS-III Off score, whereas for the Bradykinesia & Rigidity model, these were the total MDS-UPDRS-III Off score and the Total Tremor Off score.

Discussion

Using one of the largest clinical STN-DBS datasets to date, we developed ML regression models capable of predicting four primary motor outcomes following STN-DBS treatment in PD: Total MDS-UPDRS-III score and subscores for Tremor, Axial, and Bradykinesia & Rigidity. We adopted a truly predictive approach by using preoperative data to estimate actual postoperative outcomes, and validated our models with unseen test data. Our models' ability to cope with the varied responses to DBS treatment is evident from the low RMSE values indicating the mean error in predicted scores (9.1 for Total MDS-UPDRS-



(caption on next page)

Fig. 3. Model predictions and learning curves for each regression model. Panels A, C, E, G: Scatter plots contrasting predicted versus true MDS-UPDRS-III subscores (Total, Tremor, Axial, and Bradykinesia + Rigidity, respectively). Each plot features a dark line symbolizing ideal predictions and dashed gray lines denoting the variance in these predictions. Panels B, D, F, H: Learning curves for the optimal regressor, depicted for each MDS-UPDRS-III subscore category. These curves exhibit both training and test scores in relation to the number of training samples, with shaded regions indicating the standard deviation of scores across cross-validation folds.

III model; 2.6 for Tremor; 2.5 for Axial; and 5.3 for Bradykinesia & Rigidity). In contrast to previous studies using smaller cohorts or binary (responder/non-responder) outcomes, our large dataset and continuous regression framework enable the prediction of absolute postoperative scores. Moreover, contrary to methodologies that categorize outcomes in terms of absolute or relative improvement [6,31], our approach focused on predicting the absolute outcome scores after surgery in order to avoid contaminating the outcome with preoperative data, which may induce biases. Overall, by steering away from relative or threshold-based improvements, we aim to provide more personalised guidance to clinicians and patients, thus offering a nuanced addition to the existing literature. We focused on absolute postoperative Off-med, On-DBS MDS-UPDRS-III scores because they map directly onto the clinical conversation about expected status after surgery, rather than relative change metrics that can be distorted by baseline severity. In practice, interpretation often uses context-specific cut-offs or minimal clinically important differences [41]. Accordingly, continuous predictions should be complemented by such clinically meaningful thresholds to support day-to-day decision-making.

Although daily care reflects the On-med/On-DBS state, we used the Off-med/On-DBS state because it provides a standardized, medication-independent readout of stimulation effect. For clinical counseling, absolute Off-med/On-DBS predictions can be paired with the patient's typical dopaminergic response to approximate expected On-med/On-DBS status.

By design, this endpoint captures a composite of baseline severity, interval disease progression, and DBS effect rather than a pure treatment effect; therefore, we modeled the follow-up interval ("days between screening and follow-up") and examined its influence during feature selection. Consistent with this, the interval was not retained by RFECV for any outcome, suggesting limited incremental value in this cohort while acknowledging that the MDS-UPDRS is also a longitudinal progression indicator in Parkinson's disease [47]. Prospective or external validation will be needed to further disentangle progression from stimulation effects.

As shown in Fig. 3 (Panels A, C, E, and G), the models accurately predict outcomes for the majority of patients. To interpret the performance of the models, it is convenient to accompany RMSE values with a measure of the total variance in the sample, therefore we also computed R-squared values. However, the R-squared metric alone does not provide information on the predictive power of the model on unseen data. The R-squared values we obtained, ranging between 0.25 and 0.31 for the different models, suggested that the models captured some variability in the data but did not fully explain the dependent variable. These values are in line with those reported in the literature using regular linear regression on improvement scores as outcome variable.

For example, a study with 33 patients reported an R-squared of 0.38 for postoperative motor symptom improvement (percent change in motor functioning from the baseline Off state to the postoperative On-On state as measured by the MDS-UPDRS-III) based on dispositional optimism (Life Orientation Test-Revised questionnaire), depressive symptoms (Geriatric Depression Scale-Short Form), anxiety symptoms (State-Trait Anxiety Inventory), age at baseline and disease duration [27]. Another study with 181 patients, using a linear regression using sex, disease duration, cognitive status, and motor and axial levodopa response explained 12.9 % of the variance in MDS-UPDRS III motor improvement and 10.6 % in daily Off time (MDS-UPDRS IV item 3), while it explained only 3.0 % for dyskinesia (MDS-UPDRS IV item 1) [42]. In the same cohort, a multivariable model, including pre-operative levodopa response, explained 21.2 % of the variance in daily Off-time reduction [43]. A retrospective study with 61 patients using regularized linear regression reported an R-squared of 0.29 and RMSE of 14.9 for motor improvement using clinical predictors such as levodopa response, levodopa equivalent daily dose, age, sex, handedness, and preoperative MDS-UPDRS-III [44]. We emphasize that these studies evaluated relative improvement as the outcome variable and that our R-squared values were determined for out-of-sample data, which complicates a direct comparison. Of note is also that our models explain considerably more variability than the correlations between symptom severity and local field potential measures from the STN estimated at around an R-squared of 0.17 [45]. In this context, preoperative clinical parameters alone do seem to hold substantial information on motor outcomes on DBS despite measurement errors during the MDS-UPDRS screening process and other unexplained variability. We acknowledge that traditional regression or correlation analyses can identify gross predictors of outcome. However, ML methods better account for non-linear effects and complex feature interactions, thus potentially capturing subtleties in STN-DBS responsiveness.

After a comprehensive methodology encompassing feature selection, outlier removal, training and hyperparameter optimization, the learning curves analysis indicated that adding more data to the training set may not significantly improve the models' performance. Instead, the plateau observed in the curves suggests that the key to enhancing the models' accuracy lies in identifying novel predictors. Future efforts should focus on exploring new and potentially relevant factors and incorporating them into the model development process. Characteristics of structural and functional preoperative MRI scans may be suitable for this [31]. Moreover, lead placement and programming parameters have a determinant influence on DBS outcome [46]. While this study focuses on outcomes as measured by MDS-UPDRS-III, future research should incorporate broader outcome measures, such as quality of life and cognitive function, to provide a more integral view of DBS efficacy.

SHAP-based explanations in our study should be interpreted cautiously. They summarize associations learned by the fitted model and are not causal. Correlated predictors can share or exchange attribution, and local explanations vary across patients. We therefore use SHAP primarily to communicate which preoperative variables most influenced predictions and to generate hypotheses, not to guide programming or patient selection. Clinical use of such explanations should follow external validation and calibration within each center. Across all four models, higher baseline motor severity (e.g., MDS-UPDRS-III OFF/ON and subscore domains) was associated with higher predicted postoperative scores, whereas greater dopaminergic responsiveness (% improvement) was associated with lower predicted scores. Age showed comparatively smaller effects.

It should be noted that the dataset used in this retrospective study only included those patients who have already been deemed suitable candidates for STN-DBS. Thus, the regression predictive models reported here are only suitable to estimate treatment outcome among the specific group of PD patients that have already successfully undergone the initial screening process. Given the datasets focus on patients already deemed suitable for STN-DBS, our models are not designed to assist in the primary selection of candidates. This limitation reflects the need for future studies to include data from a broader range of patients, including those ultimately not selected for DBS. As well, a key limitation is the absence of an independent external cohort; future work will seek multicentre validation.

Moreover, these models were trained with data that contains inherent error. The MDS-UPDRS is sensitive to the biases and errors that clinicians make during the screening process. As such, the reliability of predictive models using these scores could be compromised. Despite these challenges, our models demonstrate promising ability to predict treatment outcomes. More accurate data collection to reduce noisy data could further enhance the models' performance by reducing outliers, thereby increasing their predictive power and reliability for clinical use. However, while the predictive power of the models is promising, they should be used in conjunction with other clinical assessments rather than as a standalone tool.

Author contributions

LB designed the research question. Alongside **JF, LB** developed the computational models, curated the data, performed the analysis, and prepared the manuscript and figures. **RdB, BvW, RS, AH**, and **JG** provided specialist consultancy and contributed to manuscript writing and revisions. **MB** led the data collection, provided the data, offered strategic guidance to address the research question, and contributed to the writing and reviewing of the manuscript.

Funding sources for study

This is an EU Joint Programme – Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organizations under the aegis of JPND – www.jpnd.eu: the Netherlands Organisation for Health Research and Development (ZonMw) – The Netherlands, and the Luxembourg National Research Fund.

Declaration of Competing Interest

This is an EU Joint Programme – Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organizations under the aegis of JPND – www.jpnd.eu: the Netherlands Organisation for Health Research and Development (ZonMw) – The Netherlands. Rob de Bie received funding for research from ZonMw, Stichting Parkinson Fonds, Parkinson Vereniging, Stichting Bavalo, Stichting Universitas, Romo foundation, and Medtronic and Bial, all paid to the institution.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jdbs.2025.10.002.

References

[1] Limousin P, Foltynie T. Long-term outcomes of deep brain stimulation in Parkinson disease. Nat Rev Neurol 2019;15(4):234–42. https://doi.org/10.1038/s41582-019-0145-9.

[2] Boutet A, Madhavan R, Elias GJB, et al. Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. Nat Commun 2021;12. https://doi.org/10.1038/s41467-021-23311-9.

[3] Kringelbach ML, Jenkinson N, Owen SLF, Aziz TZ. Translational principles of deep brain stimulation. Nat Rev Neurosci 2007;8(8):623–35. https://doi.org/10.1038/nrn2196.

[4] Perlmutter JS, Mink JW. Deep brain stimulation. Annu Rev Neurosci 2006;29:229–57. https://doi.org/10.1146/annurev.neuro.29.051605.112824.

[5] Beudel M, Oterdoom DLM, van Egmond ME, van Laar T, de Koning-Tijssen MAJ, van Dijk JMC. Diepe hersenstimulatie voor bewegingsstoornissen. Ned Tijdschr Geneeskd 2019;163(34).

[6] Habets JGV, Janssen MLF, Duits AA, et al. Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease—proof of principle in a retrospective cohort. PeerJ 2020;8:e10317. https://doi.org/10.7717/peerj.10317.

[7] Hariz M. My 25 stimulating years with DBS in Parkinson's disease. J Park Dis 2017;7(s1). https://doi.org/10.3233/JPD-179007.

[8] Hariz MI. Complications of deep brain stimulation surgery. Mov Disord 2002;17(S3):S162–6. https://doi.org/10.1002/mds.10159.

[9] Charles PD, Van Blercom N, Krack P, et al. Predictors of effective bilateral subthalamic nucleus stimulation for PD. Neurology 2002;59(6):932–4. https://doi.org/10.1212/WNL.59.6.932.

[10] Schüpbach M, Rau J, Knudsen K, et al. Neurostimulation for Parkinson's disease with early motor complications. N Engl J Med 2013;368(7):610–22. https://doi.org/10.1056/NEJMoa1205158.

[11] Rodriguez-Oroz MC, Obeso JA, Lang AE, Houeto JL, Pollak P, Rehncrona S, et al. Bilateral subthalamic stimulation in Parkinson's disease: a multicentre study with 4 years follow-up. Brain 2005;128:2240–9. https://doi.org/10.1093/brain/awh571.

[12] Brown RG, Marsden CD. Cognitive function in Parkinson's disease: from description to theory. Trends Neurosci 1990;13(1):21–9. https://doi.org/10.1016/0166-2236(90)90058-I.

[13] Dubois B, Pillon B. Cognitive deficits in parkinson's disease. J Neurol 1996;244(1):2–8. https://doi.org/10.1007/PL00007725.

[14] Lachenmayer ML, Mürset M, Antih N, et al. Subthalamic and pallidal deep brain stimulation for Parkinson's disease—meta-analysis of outcomes. npj Park Dis 2021;7:77. https://doi.org/10.1038/s41531-021-00223-5.

[15] Cernera S, Eisinger RS, Wong JK, et al. Long-term Parkinson's disease quality of life after staged DBS: STN vs GPi and first vs second lead. npj Park Dis 2020;6:13. https://doi.org/10.1038/s41531-020-0115-3.

[16] Derost PP, Ouchchane L, Morand D, et al. Is DBS-STN appropriate to treat severe Parkinson disease in an elderly population? Neurology 2007;68(17):1345–55. https://doi.org/10.1212/01.wnl.0000260059.77107.c2.

[17] Bourne SK, Conrad A, Konrad PE, Neimat JS, Davis TL. Ventricular width and complicated recovery following deep brain stimulation surgery. Stereo Funct Neurosurg 2012;90:167–72. https://doi.org/10.1159/000338094.

[18] Karsten W, Christine D, Paul K, Jens V, et al. Negative impact of borderline global cognitive scores on quality of life after subthalamic nucleus stimulation in Parkinson's disease. J Neuro Sci 2011;310(1):261–6. https://doi.org/10.1016/j.jns.2011.06.028.

[19] Kleiner-Fisman G, Herzog J, Fisman DN, Tamma F, et al. Subthalamic nucleus deep brain stimulation: summary and meta-analysis of outcomes. Mov Disord 2006;21:S290–304. https://doi.org/10.1002/mds.20962.

[20] Williams A, Gill S, Varma T, Jenkinson C, et al. Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. Lancet Neurol 2010;9(6):581–91. https://doi.org/10.1016/S1474-4422(10)70093-4.

[21] Kenney L, Rohl B, Lopez FV, Lafo JA, Jacobson C, et al. The UF deep brain stimulation cognitive rating scale (DBS-CRS): clinical decision making, validity, and outcomes. Front Hum Neurosci 2020;14. https://doi.org/10.3389/fnhum.2020.578216.

[22] Tang V, Zhu XL, Lau C, et al. Pre-operative cognitive burden as predictor of motor outcome following bilateral subthalamic nucleus deep brain stimulation in Parkinson's disease. Neurol Sci 2022;43:6803–11. https://doi.org/10.1007/s10072-022-06370-8.

[23] Sheng-Tzung T, Sheng-Huang L, Yu-Cheng C, Yan-Hong P, et al. Prognostic factors of subthalamic stimulation in Parkinson's disease: a comparative study between Short- and Long-Term effects. Stereo Funct Neurosurg 2009;87(4):241–8. https://doi.org/10.1159/000225977.

[24] Welter ML, Houeto JL, Tezenas du Montcel S, Mesnage V, et al. Clinical predictive factors of subthalamic stimulation in Parkinson's disease. Brain 2002;125(3):575–83. https://doi.org/10.1093/brain/awf050.

[25] Okun MS, Wu SS, Foote KD, Bowers D, Gogna S, et al. Do stable patients with a premorbid depression history have a worse outcome after deep brain stimulation for Parkinson disease? Neurosurgery 2011;69(2):357–61. https://doi.org/10.1227/NEU.0b013e3182160456.

[26] Soulas T, Sultan S, Gurruchaga J-M, Palfi S, Fénelon G. Depression and coping as predictors of change after deep brain stimulation in Parkinson's disease. World Neurosurg 2011;75(3):525–32. https://doi.org/10.1016/j.wneu.2010.06.015.

[27] Ray H, Cook Maher A, MacKenzie W, Zeitlin L, Chou KL, et al. The impact of dispositional optimism and depression on post-operative motor functioning following deep brain stimulation surgery for Parkinson's disease. Park Relat Disord 2020;81:41–4. https://doi.org/10.1016/j.parkreldis.2020.10.012.

[28] Deo RC. Machine learning in Medicine. Circulation 2015;132(20):1920–30. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.

[29] Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical Medicine. N Engl J Med 2016;375(13):1216–9.

[30] Habets JGV, Herff C, et al. Multicenter validation of individual preoperative motor outcome prediction for deep brain stimulation in Parkinson's disease. Stereo Funct Neurosurg 2022;100(2):121–9. https://doi.org/10.1159/000519960.

[31] Liu Y, Xiao B, Zhang C, et al. Predicting motor outcome of subthalamic nucleus deep brain stimulation for Parkinson's disease using quantitative susceptibility mapping and radiomics: a pilot study. Front Neurosci 2021;15:731109. https://doi.org/10.3389/fnins.2021.731109.

[32] Cerasa A. Machine learning on Parkinson's disease? Let's translate into clinical practice. J Neurosci Methods 2016;266:161–2. https://doi.org/10.1016/j.jneumeth.2015.12.005.

[33] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825–30.

[34] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the thirty first international conference on neural information processing systems 2017:4768–77.

[35] McKinney W. Data structures for statistical computing in python. In: Proceedings of the ninth Python in science conference 2010;445:51–6.

[36] Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.

[37] Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3): 90–5. https://doi.org/10.1109/MCSE.2007.55.

[38] Akiba T, Sano S, Yanase T, et al. Optuna: a Next-generation hyperparameter optimization framework. In: Proceedings of the twebty fifth ACM SIGKDD international conference on knowledge discovery & data mining (KDD '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 2623–31. https://doi.org/10.1145/3292500.3330701.

[39] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017;18 (17):1–5.

[40] Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. Technometrics 1999;41(3):212–23. https://doi.org/10.1080/00401706.1999.10485670.

[41] Lim SY, Tan AH. Historical perspective: the pros and cons of conventional outcome measures in Parkinson's disease. Park Relat Disord 2018;46:S47–52. https://doi.org/10.1016/j.parkreldis.2017.07.029.

[42] Artusi CA, Ledda C, Rinaldi D, Montanaro E, et al. Axial symptoms as main predictors of short-term subthalamic stimulation outcome in Parkinson's disease. J Neurol Sci 2023;453:120818. https://doi.org/10.1016/j.jns.2023.120818.

[43] Ledda C, Artusi CA, Imbalzano G, Bozzali M, et al. Predictive factors for a favorable STN-DBS outcome in Parkinson's disease: the role of axial symptoms [Accessed 12 June 2024], Mov Disord 2024;37:2022. ⟨https://www.mdsabstracts.org/abstract/predictive-factors-for-a-favorable-stn-dbs-outcome-in-parkinsons-disease-the-role-of-axial-symptoms/⟩. Accessed June 12,.

[44] Younce J, Norris S, Perlmutter J. Quantifying the value of multimodal MRI in outcomes prediction for STN DBS in PD [abstract]. Mov Disord 2023;38. ⟨https://www.mdsabstracts.org/abstract/quantifying-the-value-of-multimodal-mri-in-outcomes-prediction-for-stn-dbs-in-pd/⟩. Accessed June 12, 2024.

[45] van Wijk BCM, de Bie RMA, Beudel M. A systematic review of local field potential physiomarkers in Parkinson's disease: from clinical correlations to adaptive deep brain stimulation algorithms. J Neurol 2023;270(2):1162–77. https://doi.org/10.1007/s00415-022-11388-1.

[46] Bot M, Schuurman PR, Odekerken VJJ, et al. Deep brain stimulation for Parkinson's disease: defining the optimal location within the subthalamic nucleus. J Neurol Neurosurg Psychiatry 2018;89(5):493–8. https://doi.org/10.1136/jnnp-2017-316907.

[47] Bartl M, Dakna M, Schade S, et al. Longitudinal change and progression indicators using the movement disorder Society-Unified Parkinson's disease rating scale in two independent cohorts with early Parkinson's disease. J Park's Dis 2021;12(1): 437–52. https://doi.org/10.3233/JPD-212860.