PassFlow: a multimodal workflow for predicting deep brain stimulation outcomes

Maxime Peralta, Claire Haegelen, Pierre Jannin, John S H Baxter

▶ **To cite this version:**

Maxime Peralta, Claire Haegelen, Pierre Jannin, John S H Baxter. PassFlow: a multimodal workflow for predicting deep brain stimulation outcomes. International Journal of Computer Assisted Radiology and Surgery, 2021, 16 (8), pp.1361-1370. 10.1007/s11548-021-02435-9 . hal-03225637v2

HAL Id: hal-03225637
https://hal.science/hal-03225637v2

Submitted on 2 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PASSFLOW: A MULTIMODAL WORKFLOW FOR PREDICTING DEEP BRAIN STIMULATION OUTCOMES

Maxime Peralta
Univ Rennes
Inserm, LTSI - UMR 1099
F-35000 Rennes, France

Claire Haegelen
Univ Rennes, CHU Rennes
Inserm, LTSI - UMR 1099
F-35000 Rennes, France

Pierre Jannin
Univ Rennes
Inserm, LTSI - UMR 1099
F-35000 Rennes, France
pierre.jannin@univ-rennes1.fr

John S.H. Baxter
Univ Rennes
Inserm, LTSI - UMR 1099
F-35000 Rennes, France
jbaxter@univ-rennes1.fr

August 2, 2021

ABSTRACT

Purpose: Deep Brain Stimulation (DBS) is a proven therapy for Parkinson's Disease (PD), frequently resulting in an enhancement of motor function. Nonetheless, several undesirable side effects can occur after DBS, which can worsen the quality of life of the patient. Thus, the clinical team has to carefully select patients on whom to perform DBS. Over the past decade, there have been some attempts to relate pre-operative data and DBS clinical outcomes, with most focused on the motor symptomatology. In this paper, we propose a machine learning based method able to predict a large number of DBS clinical outcomes for PD. **Methods:** We propose a multimodal pipeline, referred to as PassFlow, which predicts 84 clinical post-operative clinical scores. PassFlow is composed of an artificial neural network to compress clinical information, an image processing method from the state-of-the-art to extract morphological biomarkers our of T1 imaging, and an SVM to perform the regressions. We validated PassFlow on 196 PD patients who undergone a DBS. **Results:** PassFlow showed correlation coefficients as high as 0.71 and were able to significantly predict 63 out of the 84 scores, outperforming a comparative linear method. The number of metrics that are predicted with this pre-operative information was also found to be correlated with the number of patients with this information available, indicating that the PassFlow method is still actively learning. **Conclusion:** We presented a novel, machine learning based pipeline to predict a variety of post-operative clinical outcomes of DBS for PD patients. PassFlow took into account various bio-markers, arising from different data modalities, showing high correlation coefficients for some scores from pre-operative data only. It indicates that many clinical outcomes of DBS can be predicted agnostic to the specific simulation parameters, as PassFlow has been validated without such stimulation-related information.

***K*eywords** Deep brain stimulation · clinical prediction · machine learning · Parkinson's disease

1 Introduction

Parkinson's Disease (PD) is the second most common neurodegenerative disease, affecting 1% of the population over 60 years old [1]. The causes are largely unknown with multiple symptoms, and there is no way to halt the progression of the disease. However, there are therapies which enhance the quality of life of the patient, notably by alleviating motor symptomatology. Even if PD is considered primarily a motor disease, the patient is also impacted by other types of symptoms, such as dementia, depression or apathy [2]. A now common therapy for PD is Deep Brain

Stimulation (DBS), a neurosurgical procedure consisting in implanting one or two electrodes in order to electrically stimulate deep anatomical structures. The most common targets are the Sub-Thalamic Nucleus (STN), the Globus Pallidus internus (GPi) and the Thalamic Ventral Intermediate Nucleus (VIM). One reason why DBS is an increasingly common therapy is that it often gives better and more stable motor outcomes than pharmaceutical therapies alone. However, as with every surgery, the procedure has potential risks such as intracranial hemorrhage [3]. Moreover, stimulation of adjacent structures and white matter tracts can possibly aggravate other symptoms. The literature has, for example, identifies declined verbal fluency [4, 5] or loss of verbal memory [5] as possible side-effects. This overall neuropsychological and cognitive impairment can lower the patient's quality of life [2]. The heterogeneity of PD means that the tasks of monitoring and treating appropriately an individual patient are complex [6]. Altogether, these factors make patient selection crucial, especially in the STN [7]. To date, it is still quite difficult for the clinician to anticipate the post-operative effects of DBS. Even if major guidelines have been defined [8], there is no way to know exactly what the positive and negative effects of DBS for a particular patient could be and there is a pressing need for addressing this problem [9].

Many pre-operative modalities are know to be predictive of the effect of DBS. Demographics have been extensively considered and age [10][11] and disease duration [10] are both predictive factors for motor improvement and for post-operative complications. In terms of imaging, T2 relaxation time in the STN is related to motor improvement [12][13] and gray matter density maps have been shown to predict weak medication responders [14]. Peralta *et al.* [15] have also shown that differences in the shape of the striatum also reflects differences in the underlying stage and type of Parkinson's disease.

Recent attempts have also combined information from several modalities. Habets *et al.* [16] used logistic regression to successfully predict weak motor responders for STN-DBS from pre-operative demographics and clinical tests, such as motor and neuropsychological assessments. Frizon *et al.* [17] also used logistic regression to predict post-operative quality of life from predictors arising from various data modalities, including demographics, clinical tests and imaging biomarkers. Shamir *et al.* [18] developed a linear function linking several multi-modal variables, such as age at surgery, Volume of Tissue Activated (VTA) within the target, and levodopa response to the post-operative UPDRS3 score, achieving satisfactory performance. These attempts demonstrate to us that the clinical outcomes of DBS are highly multi-factorial, and that information can lie in multiple data modalities. An efficient pre-operative predictive workflow should thus be able to take into account several modalities simultaneously, in a simple and robust manner. The fact that the solution lies in multiple complex and high-dimensional modalities makes this prediction task complex for a human and thus better suited for machine learning [19].

What is currently lacking in the literature is the consideration of the diversity of clinical scores collected. Indeed, a large amount of effort has been directed towards predicting the motor outcomes, often the mean of the UPDRS3 score. This is problematic in the sense that, as already stated, the success of a DBS intervention is a trade-off between desired motor enhancement and occurrence of undesirable side effects. Therefore, a pre-operative clinical outcome predictive workflow would greatly benefit from being able to predict every clinical score including neuropsychological ones [2].

In addition, the problem is often simplified though binarization, e.g. by defining a threshold between a 'good outcome' and a 'bad outcome' [14] [16] [20]. However, this approach limits the impact of such systems as the threshold should be left to the clinicians and the patients on a case-by-case basis, and not hard-coded in the predictive system. The most the predictive system should do is to predict the post-operative scores as accurately as possible, in order to provide clinicians with additional decision making support.

Contributions

In this paper, we propose a multi-modal, machine learning -based workflow, referred to as PassFlow, able to predict the clinical outcomes of DBS for PD, by using pre-operative features. We envisage that clinicians could use the system to predict the outcomes of DBS for individual patients, allowing them to better determine whether the intervention is likely to have a sufficiently positive impact on the patient's symptomatology or not. PassFlow was able to significantly predict 63 out of 82 different clinical outcomes (presented in Table 4) with correlation coefficients up to 0.71, outperforming a linear baseline. We conducted our study in respect with the applicable TRIPOD recommendations [21]. Our workflow consists of a custom supervised Artificial Neural Network (ANN), which extracts features from pre-operative clinical scores in the form of a low-dimensional vector. Then, we append additional features such as demographics, and compressed striatal shape displacements to this vector, before making the final predictions of the desired post-operative clinical score with a linear Support Vector Machine (SVM). Due to the complexity of the patient selection process and patient-specificity of the particular factors under consideration, we determine the success of our machine-learning model in terms of its ability to predict the entire corpus of post-operative clinical scores.

Table 1: Available demographics and surgical target information on the Rennes cohort.

Age at surgery	Gender (F/M)	Target (STN/GPi/VIM)
59.9 ($\pm$ 8.1)	68/94	89/54/19

Table 2: Number of complete, incomplete, and missing modalities in the dataset in terms of number of patients.

Modality	Clinical						Imag.	Add.
Phase	-6M	-3M	3M	6M	1Y	3Y		
Complete	0	23	18	1	1	0	180	133
Incomplete	150	83	81	124	103	64	0	28
Missing	46	90	97	71	92	132	16	35

2 Materials and Methods

2.1 Data

For this study, we used a cohort of PD patients who have undergone a DBS intervention at the Rennes University Hospital between May 2014 and September 2019. It consists of three data modalities: clinical, imaging, and demographic or intervention-specific.

The cohort contains 196 patients with electrodes implanted in the STN, the GPi or the VIM. Characteristics of this cohort are presented in Table 1. This cohort of patients suffers from missing modalities and missing values within each modality. Indeed, for the clinical modality, data vectors may not be complete. For example, there were two pre-operative visits only for patients undergoing a STN-DBS, and only one for the other targets. Table 2 presents the distribution of totally missing, complete and incomplete modalities. Table 3 presents the percentage of patients for which clinical values are available, at different phases and for different score categories.

Clinical data Patients had one or two pre-operative visits (approximately six and three months before the surgery, respectively referred as '-6M' and '-3M'), and up to four post-operative visits (approximately three months, six months, one year and three years after the surgery, respectively noted '3M', '6M', '1Y' and '3Y'). We did not include later visits due to the low number of datapoints. Table 4 shows the clinical tests performed at each pre-operative and post-operative visit.

Imaging data Each patient had one preoperative 3T T1- and T2-weighted MRI scan (1mm iso., Philips Medical Systems). The T2-weighted images were not used in this study.

Demographic/intervention-specific data As additional information, we collected the patients' gender, age at surgery, stimulation target (STN, GPi, VIM) and laterality (left, right or bilateral).

2.2 Proposed Method

The proposed method, referred to as PassFlow (for Patient Screening Support workFlow), consists of an SVM, with a linear kernel, which receives as input: a compressed clinical data vector, four compressed striatal shape displacement vectors, and demographic and intervention-specific data. This additional information has not been treated like other clinical data, since they were rarely incomplete and were primarily categorical, which our current clinical data handling method did not support.

The methods to compress the clinical data and extract the compressed striatal shape displacement vectors are presented in the following sections.

Table 3: Percentage of clinical scores available in the database, for different phases and clinical score categories.

Category	-6M	-3M	3M	6M	1Y	3Y	Total
Motor	68.9%	44.5%	41.5%	54.8%	44.8%	28.1%	47.1%
Behavioral	47.1%	38.1%	39.4%	40.1%	35.6%	24.2%	37.4%
Cognitive	39.9%	36.8%	39.7%	35.8%	33.2%	22.1%	34.6%
Total	52.3%	40%	40.3%	43.8%	38.1%	24.8%	39.9%

Table 4: Clinical tests performed at each visit of the patients. These tests can be divided in three categories: motor, behavioral and cognitive tests.

Test name	Function	Details
Apathy Evaluation Scale (AES)	Behavioral	
AMDP-AT scale	Behavioral	
Ekman 60-faces test	Cognitive	
Categorical Fluency	Cognitive	
Lexical Fluency	Cognitive	
Verbal Fluency	Cognitive	
Hoehn and Yahr (H&Y)	Motor	Performed off, and on dopa
Montgomery-Asberg Depression Rating Scale (MADRS)	Behavioral	
Mattis Dementia Rating Scale (MDRS)	Cognitive	
Schwab and England (S&E)	Motor	Performed off, and on dopa
Stroop Test	Cognitive	
Trail Making Test (TMT)	Cognitive	Scores A, B and B-A used.
UPDRS part 1	Behavioral	
UPDRS part 2	Motor	Performed off, and on dopa
UPDRS part 3	Motor	Performed off, and on dopa

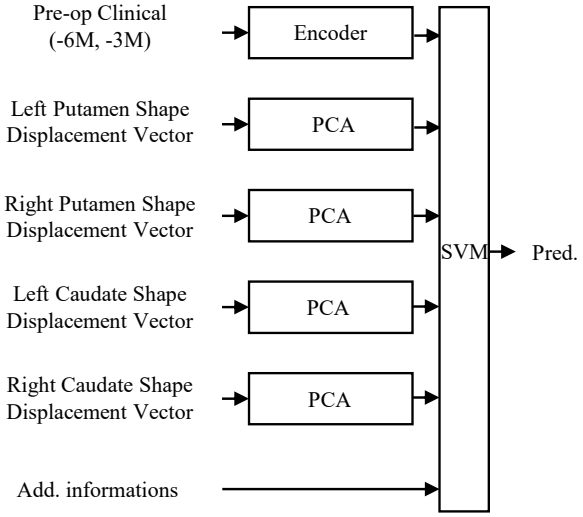


Figure 1: Schema of PassFlow.

2.2.1 Imaging data processing

Out of the T1-weighted MRI, we extracted four striatal shape displacement vectors, following the methodology presented by Khan *et al.* [22] which uses deformable registration to extract the the bilateral caudate and the bilateral putamen in the patient MRI and compare their shape against those in the MNI PD25 Atlas [23]. The shape deformation field sampled at the surface of each of the four structure were then re-organized as a 1D vectors.

As these vectors were high-dimensional (several thousand values), we compressed these vectors separately using a Principal Component Analysis (PCA). The number of principal components kept was optimized using the Hyper-Parameter Optimization (HPO) process presented in Section 2.5, for each post-operative clinical score.

We chose this methodology to extract bio-markers from T1-weighted MRI because we previously shown their interest as PD staging bio-marker, and their correlation with the patient's UPDRS3 score [15].

2.2.2 Clinical data processing

Clinical test results from normal clinical routine present two problems for downstream analysis: they are high-dimensional, and they very frequently suffer from missing data. We previously published an extensive study on

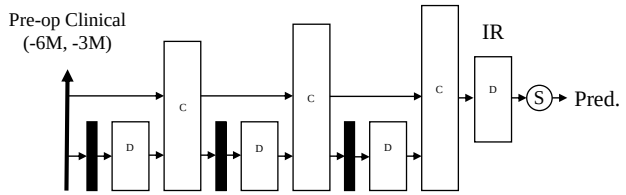


Figure 2: Schema of the ANN structure used for clinical data pre-processing. Blocks 'D' are densely connected layers with ReLU activation, blocks 'C' are concatenation layers, black-filled blocks are dropout layers with a drop rate of 0.1, neuron 'S' is a densely connected neuron with sigmoid activation.

compressing clinical data by appropriately handling missing data, and proposed a custom deeply-learned autoencoder that outperformed linear baselines such as PCA [24].

In the current study, we kept the same ANN structure, but instead of training it in an unsupervised manner, we only took the encoder part and added a sigmoid neuron after the bottleneck to perform regression, and therefore train the encoder in a supervised manner.

This ANN structure is presented in Figure 2. The input bias-only layer allows a more advanced missing data imputation strategy than regular zero imputation or mean imputation: the pre-operative clinical data vector is given as input to the ANN, as well as a mask vector indicating where the missing values are. From these two vectors, the bias only layer adds a learnable bias to the input vector at places where values are missing. This way, the optimal imputation value for each clinical score is learned in a supervised manner. Following this data imputation layer, the ANN consists of a series of dense layers followed by a dropout layer, the output of which is concatenated to the input for the successive layers.

The shape of the ANN, e.g. the depth and the number of dense neurons per layer, has been optimized with the HPO process presented in Section 2.5. The best shape was defined as the one giving the highest mean regression correlation score for all post-operative clinical scores. Therefore, the same ANN topology was used for every post-operative score.

Other hyper-parameters, such as the learning rate, the batch size and the internal representation size have been optimized for each post-operative clinical score separately.

Once the ANN has been trained in a supervised manner, we removed the sigmoid neuron and used the output of the internal representation as clinical features. One training has therefore to be done for each post-operative clinical score to predict.

2.3 Accuracy and loss metrics

We used coefficient correlation R as a comparison metric between methods, and for HPO. For the regression task, the Mean Squared Error (MSE) between the prediction and the ground-truth score was used as the ANN loss function.

2.4 Training and validation

Training and validation has been done separately for each of the 84 post-operative clinical scores we tried to predict. As our database suffers from missing values, we only evaluated PassFlow on the patients for which the post-operative clinical score, which serves as the ground-truth, is known. (That is, outcome values are not imputed as to not introduce a bias into the evaluation.) Note that the number of patients is therefore not the same for all the post-operative clinical scores as the proportion of missing data varies across the different post-operative clinical scores.

For each post-operative clinical score with more than 50 samples, we used a stratified 50-fold Cross-Validation (CV): each model has been trained 50 times separately, using one fold as a validation data and the remaining 49 as training data. If a post-operative clinical score had less than 50 samples, a Leave One Out Cross-Validation (LOOCV) has been used.

2.5 Hyper-parameter optimization

The hyper-parameters of each classifier have been optimized using Bayesian optimization with a Gaussian processes as a surrogate model and expected improvement as the criterion. The number of points tested was the square of the number of hyper-parameters to optimize plus one.

Table 5: Statistics regarding the coefficient correlations on clinical score predictions, with our proposed method PassFlow and the linear baseline.

Method	Mean R	Max R	Sig.	Not sig.	Discarded
PassFlow	$0.35(\pm 0.20)$	0.71	63	18	3
Linear	$0.27(\pm 0.23)$	0.68	44	37	

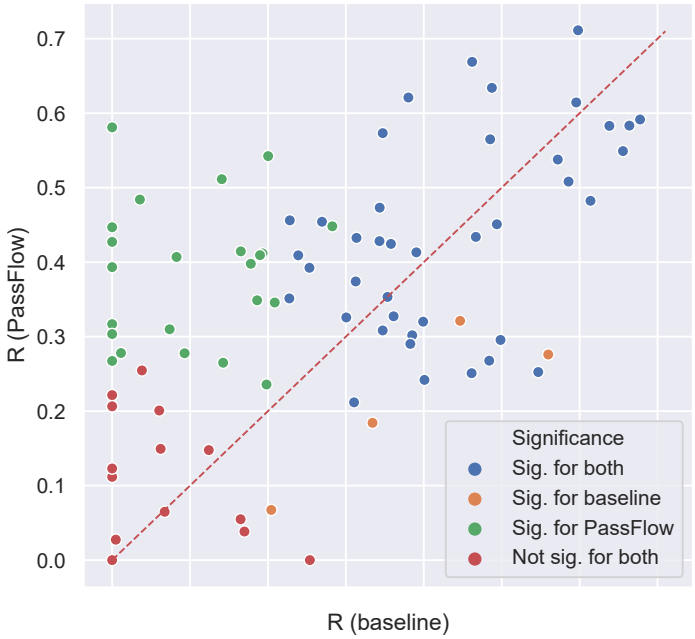


Figure 3: Comparison of the correlation coefficients (R) between the ground-truths and the predictions for PassFlow versus the linear baseline, for each the post-operative clinical scores. Points above (resp. under) the dashed red line represents scores that are predicted better by PassFlow (resp. the baseline).

We optimized the hyper-parameters controlling the shape of the neural network (depth and number of neurons per layer) on the average performances on all post-operative scores, in order to have a fixed structure for all the post-operative scores. For the hyper-parameters controlling the training process of the neural network (batch size, number of training epochs, number of principal components kept for striatal shape displacement vector, and size of the intermediate representation for clinical data), we ran a different HPO for each post-operative score. We changed the data partitioning and the random-number generator seed after HPO, in order to limit the HPO bias and report results closer to those expected from prospective use.

2.6 Software environment

We used Python 3.6, with Keras (version 2.2.4, TensorFlow version 1.12.0 as a backend) for ANN implementations, and Scikit-learn (version 0.21.1) for other machine learning implementations.

3 Experiments

We conducted an experiment, which consisted in studying the correlation coefficients (R) between the predictions and the ground truths for each post-operative clinical score of the Rennes database. Results were obtained by running a 50-fold CV for each post-operative clinical score with at least 10 values, with the optimized hyper-parameters.

3.1 Comparison against the linear baseline

Figure 3 shows the correlation coefficients of predictions for each post-operative clinical score (provided that there are more than 10 patients), comparing PassFlow with a linear baseline, which is a simple linear regression taking as input the pre-operative values (at -6M and -3M) of the clinical score being predicted.

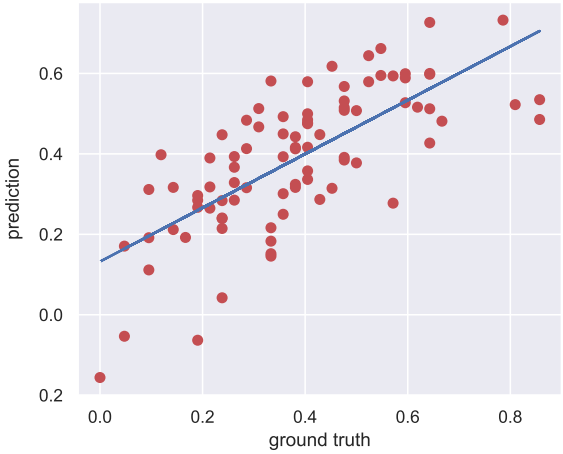


Figure 4: Scatter plots of the prediction of lexical fluencies, 3 months post-surgery, with our proposed system PassFlow. There is a correlation coefficient (R) of 0.71 between the predictions of PassFlow and the ground-truths.

Table 6: Statistics regarding the coefficient correlations on clinical score predictions, with our proposed method for the different stimulation sites represented in the cohort.

Stim. site	% of db	Mean R	Max R	Sig.	Not sig.	Discarded
GPi	33%	$0.25(\pm 0.21)$	0.77	17	63	4
STN	55%	$0.27(\pm 0.19)$	0.61	36	45	3
VIM	12%	$0.23(\pm 0.18)$	0.67	3	53	28

Concerning the 42 clinical scores that are predicted significantly by both methods, we observed that the linear baseline outperformed PassFlow for 22 of them, and the average number of patients for these clinical scores was 73.5. On the remaining 20 clinical scores, PassFlow outperformed the linear baseline, and the average number of patients for these scores was 84.3. Figure 3 indicates that these scores all lie relatively close to the dotted red line indicating equivalent overall performance.

Table 5 synthesizes the results on all scores, showing the number of tests for which we have statistically significant results and statistics on R obtained, for both PassFlow and the linear baseline. Clinical scores with less than 10 patients were discarded.

3.2 Detailed PassFlow results

Figure 5, shows the correlation coefficient between the predictions of PassFlow and the ground truths, for each post-operative clinical scores. Figure 5a shows the motor scores, Figure 5c shows the cognitive scores and Figure 5b shows the behavioral scores.

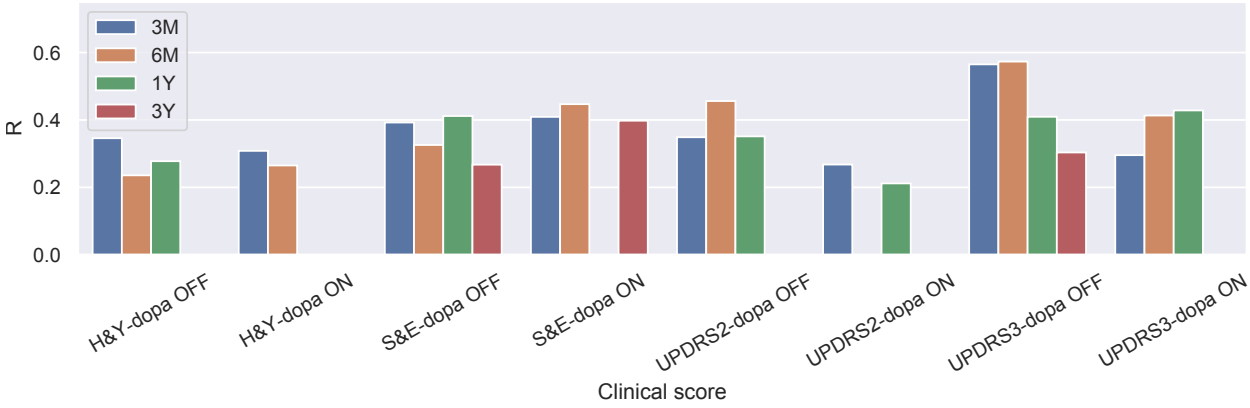
As an example, Figure 4 shows the predictions against the ground truths for the score the most accurately predicted by PassFlow ($R = 0.71$), which is lexical fluency three months after surgery.

Figure 6 shows a positive effect of the number of patient for the performances of PassFlow, as the predictions tend more to be significant if more scores are available.

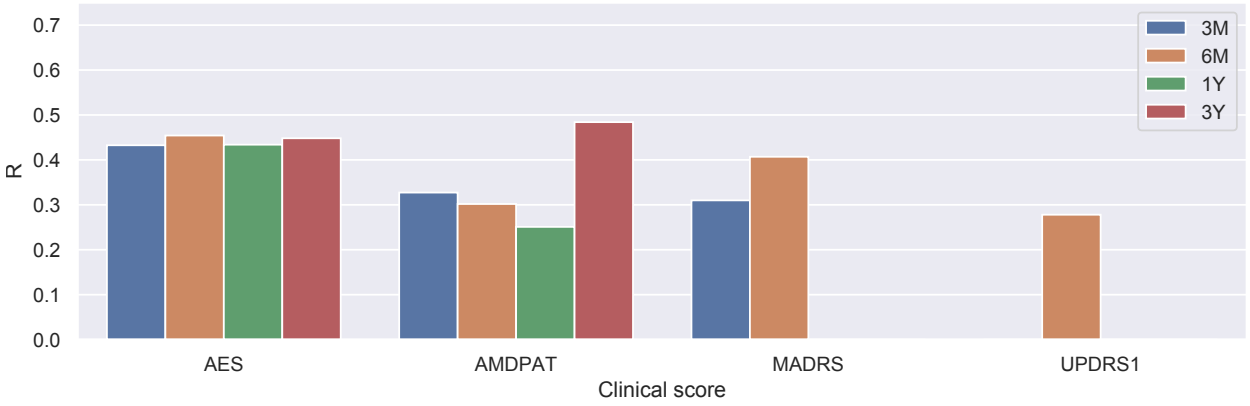
Finally, Table 6 breaks down the results of PassFlow for the different stimulation sites composing our cohort (clinical scores with less than 10 samples have been discarded). It is important to note that the number of not significant scores is high as the subdivision of the cohort drastically lowers the number of patients on which the t-test is perform, resulting in a higher R threshold above which the t-test is significant at $p < 0.05$.

4 Discussion

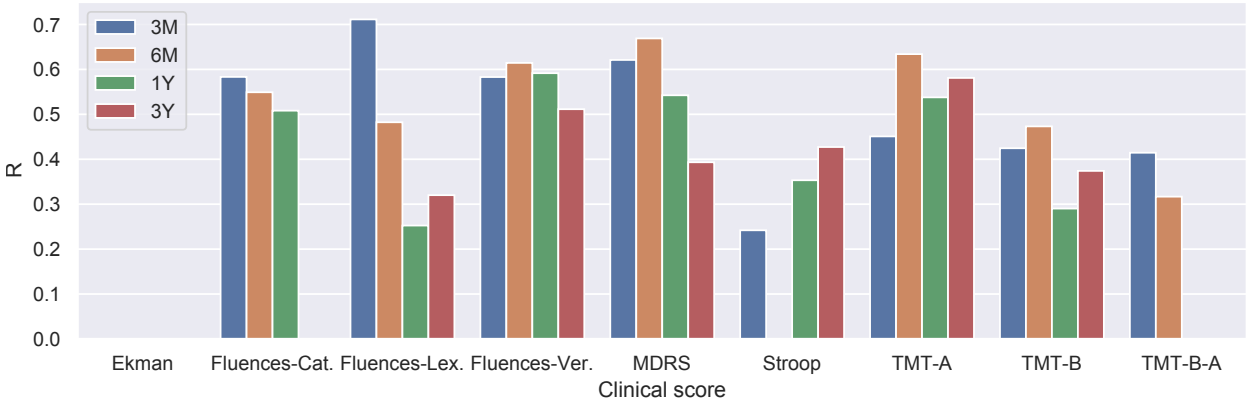
Our workflow, PassFlow, was able to significantly outperform a linear baseline Table 5 shows that PassFlow presented a higher mean and maximum correlation coefficient across all the post-operative clinical scores to predict, and was



(a) Correlation coefficient of PassFlow predictions for motor tests.



(b) Correlation coefficient of PassFlow predictions for behavioral tests.



(c) Correlation coefficient of PassFlow predictions for cognitive tests.

Figure 5: Significant correlation coefficient (R) of PassFlow predictions for each post-operative clinical scores, split between motor scores, behavioural scores, and cognitive scores. (Scores that do not result in a significant correlation have been removed for clarity.)

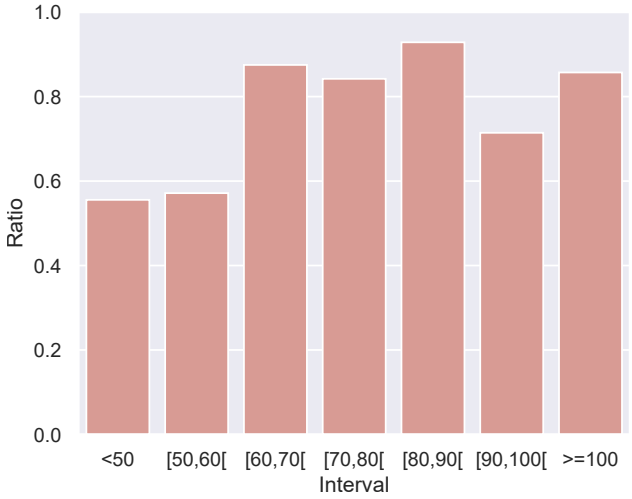


Figure 6: Ratio of significant predictions (over total predictions) for the test containing a particular number of patient scores, for PassFlow.

able to predict more clinical scores than the linear baseline. This result indicates that PassFlow successfully took advantage of the extensive pool of features used, i.e. the compressed clinical pre-operative state of the patient, imaging bio-markers and additional information such as demographics and target structure, allowing it to predict more scores than a linear method considering only one clinical score in input. Those better performances mostly come from the fact that PassFlow was able to make significant predictions for a larger number of scores than the baseline, as showed in Figure 3. Additionally, Figure 3 shows that scores predicted significantly by both methods (blue points) seem to follow a linear distribution, indicating that variability in predictions quality comes from the clinical data itself, rather than the methods.

In this extent, we showed that the number of patients is crucial in term of performances for PassFlow. We noted a clear effect of the number of patients on PassFlow's ability to make significant predictions, as showed in Figure 6. Additionally, we showed that PassFlow outperformed the baseline for clinical scores with a larger number of patients, and got outperformed for clinical scores with less patients. Both of these results showed that PassFlow can continue to learn and as room for improvement given a broader or more complete database, where it is not the case for the linear baseline. It is easily explained by the curse of dimensionality, as PassFlow took more features as input, increasing the risk of overfitting for small databases.

To go further, clinical scores at later visits seem to be harder to predict, as showed in Figure 5. This can be explained either by the fact that there is a greater variability of the disease as it progresses, or by the fact that there were fewer patients for which scores were available, rendering the training of PassFlow more difficult. Unfortunately, the high variability in the number of datasets available for each clinical score prevents us from drawing further conclusions regarding their relative complexity (that is, questionnaires undergone by a larger number of patients tended to be easier for the algorithm to predict, regardless of the questionnaire's complexity).

All together, the high number of scores successfully predicted by PassFlow indicates that some aspects of post-operative clinical effects of DBS can be anticipated pre-operatively. For certain scores, like fluencies, most of the variance seems to be explainable by pre-operative features. This may indicate that DBS surgeries are consistent in terms of electrode placements affecting these scores. Nonetheless, most of the predictions are in the low to moderate correlation range. This denotes that methodological improvements, as well as supplementary data collection efforts would be necessary to integrate PassFlow in clinic.

Longer-term outcomes appear harder to predict, which is coherent with the clinical difficulty of predicting the evolution of PD, the higher long-term effects of DBS compared to immediate results, and with the observation that our database contains longer-term outcomes data from fewer patients. Nevertheless, those longer-term outcomes are probably the most important ones to predict, as this is precisely where the clinicians and the patients could benefit from this predictive information. That being said, it seems hard to determine a threshold above which the performance of PassFlow would be considered satisfying for DBS patient selection.

Lastly, both methods are trained and evaluated with respect to a clinical ground-truth which is, by essence, inaccurate. This ground-truth is subject to the patients' own perceptions of their symptoms as well as the subjectivity of the assessor

performing the clinical evaluation. This implies that there is some operator-specific and possible patient specific biases arising from their perception of the patient's symptoms, which would not apply to a quantitative measurement of a more objective characteristic symptom. Assuming all the tests are performed by the same operator, one could imagine a machine learning algorithm being able to control for this bias, but that is often not the case. In addition, the clinical state of a patient is not stable. The patient could be having a *good* or a *bad* day when performing the test. This lack of stability could also apply to objective measurements of the patient's clinical state and would indicate an upper-bound on the possible accuracy of any predictive method. Measurements of the inter- and intra-operator variability for some of these tests, notably UPDRS, which indicate a moderate level of variability for operators [25] although, to the best of our knowledge, no such measurement has been performed with respect to inter- and intra-subject variability.

4.1 Future Work

Several aspects of this study could be extended. Firstly, relative importance of each modality could be extensively studied, globally and for each score. Given the nature of the methods and the relatively small number of datasets available for training, it is possible that having more data modalities available would actually have a negative effect on the performance due to overfitting. If the most informative data modalities could be identified, this could encourage better benchmarking of these forms of data, especially for imaging biomarkers. Secondly, if a larger amount of data becomes available, new biomarkers could be identified, investigated, and compared. There is a distinct possibility, as shown from the heterogeneity of imaging biomarkers investigated in the literature, that addition imaging information could improve predictive performance. In addition, the performance of PassFlow is very heterogeneous across the different post-operative clinical scores, and there are scores that PassFlow fails to predict. This would seem to imply that those scores, by their nature, are more difficult to predict than others. However, additional investigation would be necessary to determine the underlying reason behind this difficulty as there are several hypotheses, including the particular sensitivity of these scores to the stimulation location and parameters or increased variability due to their subjective nature. Finally, the current method has been validated on PD, but it could be extended to other conditions which are treated with DBS, provided a large enough database is available. It could also allow for comparing between predictive models, in order to see if there are common risk-factors between the diseases, and universal DBS counter indications.

4.2 Towards clinical integration

There remains a large amount of future work to perform prior to the utilisation of such a system in clinic, as several non-methodological considerations remain to be studied [26][27]. These considerations include determining what contexts such a predictive system would best aid the clinical design-making process. A specific example of this is determining which scores are most important to reconstruct for which patients, as the current model weights all outcome values equally regardless of their importance to the patient-specific decision-making process. Ultimately, decision analytic measures such as net benefit [28], to determine if using a predictive system such as PassFlow would be beneficial.

5 Conclusions

In this paper, we presented a novel, machine learning based pipeline, referred as PassFlow, to predict a variety of post-operative clinical outcomes of DBS for PD patients. PassFlow took into account various bio-markers, arising from different data modalities, such as compressed pre-operative clinical features and compressed striatal displacement. PassFlow has been able to significantly predict most post-operative clinical scores, showing high correlation coefficients for some scores from pre-operative data only, and that better performance could be achieved using larger databases. It indicates that many clinical outcomes of DBS could be predicted pre-operatively, as PassFlow has been validated without stimulation-related information. Finally, PassFlow is flexible, and can be extended to other data modalities, such as other imaging sequences or pipelines. Taken together, PassFlow represents a promising step towards computer-assisted patient screening, reducing the amount of uncertainty clinical teams have in deciding which patients could benefit from DBS.

Declarations

Funding: This work was funded by the Fondation pour la Recherche Médicale.
Conflicts of interest: The authors have no relevant financial or non-financial interests to disclose.
Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its

later amendments or comparable ethical standards.
Informed consent: Informed consent was obtained from all individual participants included in the study.
Availability of data and material: Data is not available for this study.
Code availability: Code is currently not made available.

References

[1] Garrett E Alexander. Biology of parkinson's disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder. *Dialogues in clinical neuroscience*, 6(3):259, 2004.

[2] Cynthia S Kubu. The role of a neuropsychologist on a movement disorders deep brain stimulation team. *Archives of Clinical Neuropsychology*, 33(3):365–374, 2018.

[3] Galit Kleiner-Fisman, Jan Herzog, David N Fisman, Filippo Tamma, Kelly E Lyons, Rajesh Pahwa, Anthony E Lang, and Günther Deuschl. Subthalamic nucleus deep brain stimulation: summary and meta-analysis of outcomes. *Movement disorders: official journal of the Movement Disorder Society*, 21(S14):S290–S304, 2006.

[4] Alexander I Tröster, Joseph Jankovic, Michele Tagliati, DeLea Peichel, and Michael S Okun. Neuropsychological outcomes from constant current deep brain stimulation for parkinson's disease. *Movement Disorders*, 32(3):433–440, 2017.

[5] Jae-Hyeok Heo, Kyoung-Min Lee, Sun Ha Paek, Min-Jeong Kim, Jee-Young Lee, Ji-Young Kim, Soo-Young Cho, Yong Hoon Lim, Mi-Ryoung Kim, Soo Yeon Jeong, and Beom S. Jeon. The effects of bilateral subthalamic nucleus deep brain stimulation (stn dbs) on cognition in parkinson disease. *Journal of the neurological sciences*, 273(1-2):19–24, 2008.

[6] Luke Mugge, Brianna Krafcik, Gregory Pontasch, Ahmed Alnemari, Joseph Neimat, and Daniel Gaudin. A review of biomarkers use in parkinson with deep brain stimulation: a successful past promising a bright future. *World Neurosurgery*, 123:197–207, 2019.

[7] Anthony E Lang, Jean-Luc Houeto, Paul Krack, Cynthia Kubu, Kelly E Lyons, Elena Moro, William Ondo, Rajesh Pahwa, Werner Poewe, Alexander I Tröster, Ryan Uitti, and Valerie Voon. Deep brain stimulation: preoperative issues. *Movement disorders: official journal of the Movement Disorder Society*, 21(S14):S171–S196, 2006.

[8] Pierre Pollak. Deep brain stimulation for parkinson's disease–patient selection. In *Handbook of clinical neurology*, volume 116, pages 97–105. Elsevier, 2013.

[9] Mario Giorgio Rizzone, Tiziana Martone, Roberta Balestrino, and Leonardo Lopiano. Genetic background and outcome of deep brain stimulation in parkinson's disease. *Parkinsonism & related disorders*, 64:8–19, 2019.

[10] Jurg L Jaggi, Atsushi Umemura, Howard I Hurtig, Andrew D Siderowf, Amy Colcher, Matthew B Stern, and Gordon H Baltuch. Bilateral stimulation of the subthalamic nucleus in parkinson's disease: surgical efficacy and prediction of outcome. *Stereotactic and functional neurosurgery*, 82(2-3):104–114, 2004.

[11] Farrokh Farrokhi, Quinlan D Buchlak, Matt Sikora, Nazanin Esmaili, Maria Marsans, Pamela McLeod, Jamie Mark, Emily Cox, Christine Bennett, and Jonathan Carlson. Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms. *World Neurosurgery*, 134:e325–e338, 2020.

[12] Shohei Watanabe, Koichi Suenaga, Asami Yamamoto, Kazuo Abe, Noriko Kotoura, Reiichi Ishikura, Shozo Hirota, and Hiroo Yoshikawa. Correlation of subthalamic nuclei t2 relaxation times with neuropsychological symptoms in patients with parkinson's disease. *Journal of the neurological sciences*, 315(1-2):96–99, 2012.

[13] Tarja Lönnfors-Weitzel, Thilo Weitzel, Johannes Slotboom, Claus Kiefer, Claudio Pollo, Michael Schüpbach, Markus Oertel, Alain Kaelin, and Roland Wiest. T2-relaxometry predicts outcome of dbs in idiopathic parkinson's disease. *NeuroImage: Clinical*, 12:832–837, 2016.

[14] Tommaso Ballarini, Karsten Mueller, Franziska Albrecht, Filip Růžička, Ondrej Bezdicek, Evžen Růžička, Jan Roth, Josef Vymazal, Robert Jech, and Matthias L Schroeter. Regional gray matter changes and age predict individual treatment response in parkinson's disease. *NeuroImage: Clinical*, 21:101636, 2019.

[15] Maxime Peralta, John S. H. Baxter, Ali R Khan, Claire Haegelen, and Pierre Jannin. Striatal shape alteration as a staging biomarker for parkinson's disease. *NeuroImage: Clinical*, 27, 2020.

[16] Jeroen GV Habets, Marcus LF Janssen, Annelien A Duits, Laura CJ Sijben, Anne EP Mulders, Bianca De Greef, Yasin Temel, Mark L Kuijf, Pieter L Kubben, and Christian Herff. Machine learning prediction of motor response after deep brain stimulation in parkinson's disease—proof of principle in a retrospective cohort. *PeerJ*, 8:e10317, 2020.

[17] Leonardo A Frizon, Olivia Hogue, Rebecca Achey, Darlene P Floden, Sean Nagel, Andre G Machado, and Darlene A Lobel. Quality of life improvement following deep brain stimulation for parkinson disease: development of a prognostic model. *Neurosurgery*, 85(3):343–349, 2019.

[18] Reuben R Shamir, Trygve Dolber, Angela M Noecker, Anneke M Frankemolle, Benjamin L Walter, and Cameron C McIntyre. A method for predicting the outcomes of combined pharmacologic and deep brain stimulation therapy for parkinson's disease. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 188–195, 2014.

[19] Joeky T Senders, Patrick C Staples, Aditya V Karhade, Mark M Zaki, William B Gormley, Marike LD Broekman, Timothy R Smith, and Omar Arnaout. Machine learning and neurosurgical outcome prediction: a systematic review. *World neurosurgery*, 109:476–486, 2018.

[20] Kyriaki Kostoglou, Konstantinos P Michmizos, Pantelis Stathis, Damianos Sakas, Konstantina S Nikita, and Georgios D Mitsis. Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings. *IEEE Transactions on Biomedical Engineering*, 64(5):1123–1130, 2016.

[21] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. *Circulation*, 131(2):211–219, 2015.

[22] Ali R Khan, Nole M Hiebert, Andrew Vo, Brian T Wang, Adrian M Owen, Ken N Seergobin, and Penny A MacDonald. Biomarkers of parkinson's disease: Striatal sub-regional structural morphometry and diffusion mri. *NeuroImage: Clinical*, 21:101597, 2019.

[23] Yiming Xiao, Vladimir Fonov, M Mallar Chakravarty, Silvain Beriault, Fahd Al Subaie, Abbas Sadikot, G Bruce Pike, Gilles Bertrand, and D Louis Collins. A dataset of multi-contrast population-averaged brain mri atlases of a parkinson's disease cohort. *Data in brief*, 12:370–379, 2017.

[24] Maxime Peralta, Pierre Jannin, Claire Haegelen, and John S. H. Baxter. Data imputation and compression for parkinson's disease clinical questionnaires. *Artificial Intelligence in Medicine (in press)*, 2021.

[25] Bart Post, Maruschka P Merkus, Rob MA de Bie, Rob J de Haan, and Johannes D Speelman. Unified parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Movement disorders: official journal of the Movement Disorder Society*, 20(12):1577–1584, 2005.

[26] Bilal A Mateen, James Liley, Alastair K Denniston, Chris C Holmes, and Sebastian J Vollmer. Improving the quality of machine learning in health applications and clinical research. *Nature Machine Intelligence*, 2(10):554–556, 2020.

[27] Michael J Pencina, Benjamin A Goldstein, and Ralph B D'Agostino. Prediction models-development, evaluation, and clinical application. *The New England journal of medicine*, 382(17):1583–1586, 2020.

[28] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, 352, 2016.