POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Investigating Speech Patterns in Parkinson's Disease (PD) and Rapid Eye Movement Sleep Behavior Disorder (RBD)

Nonparametric Statistics project, AY 2023-2024

Ange Dakouri, Simone Giacomello, Marta Krychkovska, Antonio Napolitano

---

Contents

1. Introduction

Background and Motivation
Neurodegenerative disorders like Parkinson's Disease (PD) profoundly impact motor functions, including speech, one of the most complex human motor skills. Traditionally, the assessment of speech abnormalities in Parkinson's Disease (PD) has been constrained by manual evaluations and laboratory analyses, which are limited by small sample sizes and the subjective nature of perceptual tests. This traditional approach often overlooks subliminal speech deficits, especially in the early stages of neurodegeneration, highlighting the necessity for more innovative, sensitive, and scalable diagnostic tools.

Automated Vocal Analysis: A New Frontier
The advent of automated vocal analysis marks a significant advancement in the early detection and characterization of neurodegenerative patterns in speech. This approach utilizes acoustic microphone data from natural connected speech, extracting critical features that indicate respiratory deficits, dysphonia, imprecise articulation, and dysrhythmia. The potential of automated vocal analysis lies in its ability to provide a non-invasive, cost-effective, and easily deployable tool for the early diagnosis of PD and other neurodegenerative disorders.

Study Overview
Our project is inspired by a seminal study [1] that compared speech recordings from individuals with Rapid Eye Movement Sleep Behavior Disorder (RBD), newly diagnosed PD patients, and healthy controls. This study demonstrated the efficacy of automated vocal analysis in detecting subliminal parkinsonian speech deficits in RBD patients, a group at high risk for developing PD. The findings underscore the potential of automated methods in identifying early indicators of neurodegeneration, paving the way for timely interventions.

Implications and Future Directions
The implications of automated vocal analysis for the screening and diagnosis of neurodegenerative disorders are profound. By facilitating early detection of PD and other synucleinopathies, this approach could significantly contribute to the development of targeted therapies and improve the quality of life for affected individuals. Our project aims to build upon these findings by applying nonparametric statistical methods to enhance the understanding and detection of neurodegenerative patterns in speech, thereby contributing to the evolving landscape of PD diagnosis and management.

2. Dataset description

The dataset we adopted is a collection of 130 observations: 30 patients with untreated Parkinson's disease (PD), 50 subjects diagnosed with RBD, and 50 healthy subjects (HC) without a history of neurological or communication disorders (control group).
Overall, the dataset contains 65 features. These features can be grouped into six major categories.

- *Demographic information*: age and gender.
- *Clinical information*: history of Parkinson's disease in the family, duration of the disease from the first symptoms, and type of medication (antidepressants, antiparkinsonians, antipsychotics and benzodiazepines).
- *Motor impairment and disease progression scores*: PD and RBD patients were assessed using the Unified Parkinson's Disease Rating Scale (UPDRS) to determine an overall score for the disease progression. Patients with PD were also assigned a score based on the Hoehn and Yahr scale.
- *Motor ability evaluation*: for each PD and RBD patient a detailed evaluation (following the UPDRS) is provided in terms of tremor, rigidity, agility, posture, facial expression, speech, finger taps, and hand movements.
- *Speech tasks examination*: each subject was asked to perform two speech-related tasks: **reading passage** and **monologue speech**. In the first task, they were presented with a written extract to read loudly, whereas in the second they had to talk about some casual topics of their choice. Both tasks were evaluated by means of articulation, phonation, respiration, and timing. Specifically, a set of 12 variables was used to describe the performance in each task.

Articulation	DUS	Duration of unvoiced stops
	DUF	Decay of unvoiced fricatives
Phonation	DVI	Duration of voiced intervals
	GVI	Gaping in-between voiced intervals
Respiration	RLR	Relative loudness of respiration
	PIR	Pause intervals per respiration
	RSR	Rate of speech respiration
	LRE	Latency of respiratory exchange
Timing	EST	Entropy of speech timing
	RST	Rate of speech timing
	AST	Acceleration of speech timing
	DPI	Duration of pause intervals

Table 1: Variables used to describe the performance in the speech tasks

3. Workflow Description

As part of our project, we conducted an analysis using a systematic approach detailed in the following steps. The corresponding image below, illustrates these steps and provides a visual representation of the process we considered in our study. This first description is only meant to give an idea of the workflow of our project. The specific methods used to develop each task is then described extensively throughout the paper.

1. **Clinical Assessment (Panel A and B):** We first considered the feature UDPRS III* (Unified Parkinson's Disease Rating Scale), which describes the traditional way of assessing Parkinson's disorders, mostly based on motor evaluations. We set a cut-off value from this scale to classify RBD patients as either motor negatives (score ≤ 3) or motor positives (score ≥ 3). This specific choice for the cut-off value was suggested by the research team who published the paper [1]. Motor positive identifies an RBD patient at high risk of developing Parkinson's disease, according to the paper's suggestion.
2. **Speech Analysis (Panel C and D):** We then focused on the speech data measured through automated vocal analysis from the research team.
3. **Pattern Analysis (Panel E):** Among all the speech features, we extracted relevant ones with the goal of identifying distinctive patterns that separate PD speech from that of healthy speech. Here we developed the core models using the selected features.
4. **Index Test Application (Panel F):** The speech features of RBD patients were processed through the pattern recognition models developed from the PD and healthy control data. RBD patients were classified as speech positives if their patterns aligned with PD patients, or speech negatives if they aligned with healthy controls.
5. **Accuracy Evaluation:** The accuracy of the speech test was determined by comparing the classification results of the RBD patients (speech positives and negatives) with their motor status as determined by UPDRS III*, assessing the predictive capability of speech features for PD-related neurodegeneration. Specifically, a True Positive (TP) was counted when a motor positive was correctly classified as a speech positive, and a True Negative (TN) when a motor negative was correctly classified as a speech negative. False Positives (FP) occurred when motor negatives were incorrectly classified as speech positives, and False Negatives (FN) when motor positives were incorrectly classified as speech negatives. The formulas used for calculating the performance metrics are as follows:

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Sensitivity: } \frac{TP}{TP+FN} \qquad \text{Specificity: } \frac{TN}{TN+FP}$$

These measures assess the predictive capability of speech features for PD-related neurodegeneration, with accuracy indicating the overall correctness, sensitivity the correct identification of PD cases, and specificity the correct identification of non-PD cases.

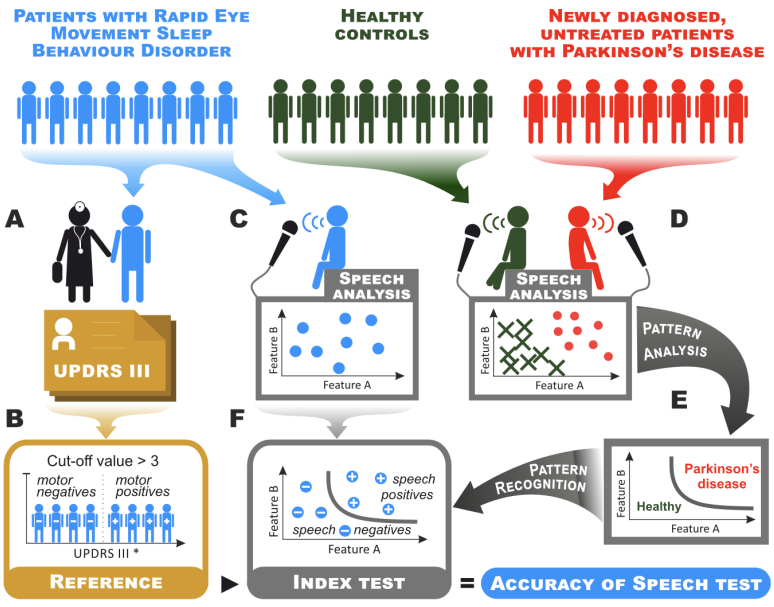


Figure 1: Visual representation of the analysis steps taken in our study.

4. Data exploration

As the first step of our analysis, we wanted to understand and investigate in detail the main characteristics of the data at hand. We noticed from the very beginning of our study that the observations are diverse and heterogeneous.

4.1. Demographic and clinical information

Starting from the demographic information (*age* and *gender*), one can observe that the dataset is not balanced. There is a significant prevalence of male subjects, especially in RBD and Control groups. From the point of view of the age distribution, we have observations ranging from 34 up to 83 years, with most of the data falling in the 50-80 range.

	F	M
Control	9	41
RBD	9	41
PD	9	21

Table 2: Gender

	31-40	41-50	51-60	61-70	71-80	81-90
Control	1	5	13	15	15	1
RBD	1	3	10	22	13	1
PD	2	1	6	11	10	0

Table 3: Age distribution

The dataset contains information on the disease duration from the first symptoms for both PD and RBD patients. While subjects affected by Parkinson's disease were all in the early stages of the disease (ongoing for a maximum of six years), subjects affected by Rapid Eye Movement sleep behavior disorder had much more variability in the duration of the disease. Some had experienced the condition for only one or two years, while others were affected for more than fifteen years.
No patients were taking antiparkinsonian or antipsychotic medication at the time when the data was recorded. However, some patients were undergoing antidepressant therapy and some were taking benzodiazepine medication (respectively 10 and 12 individuals).
Of all the subjects contained in the dataset, only three have had a positive history of Parkinson's disease in the family. Hence, the heredity/genetic factor of the disease was not taken into account in our analysis.

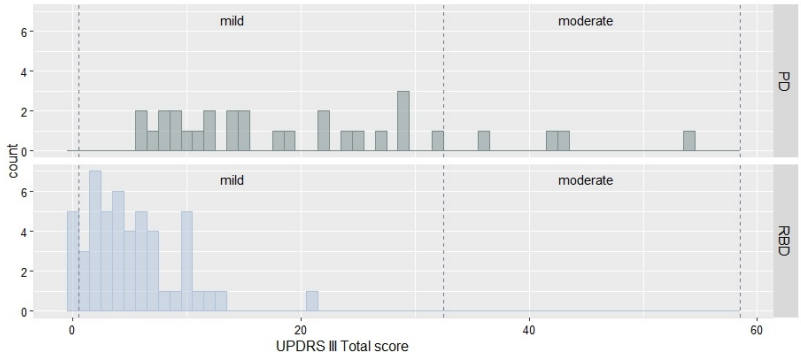
4.2. UPDRS

A classical approach to evaluate the progression of Parkinson's disease is through the Unified Parkinson's Disease Rating Scale (UPDRS). The UPDRS is the most commonly used scale in the clinical study of Parkinson's disease. This scale comprises six sections, each dedicated to a different aspect of the disease progression.
The subjects in the dataset were evaluated based on the third section of the UPDRS. This section involves a monitored motor evaluation that is scored by a clinician. It comprises 18 items, with each item having a value ranging from 0 to 4. A total motor evaluation score (*UPDRS III Total*) is obtained by summing up the scores assigned to each item, providing a value between 0 and 132.

The plot below shows the distribution of the UPDRS III Total score of PD and RBD patients contained in the dataset. We observe that most RBD patients have low scores and they all show 'mild' severity of the disease. While PD patients have higher scores, with some of them having developed 'moderate' symptoms.

0	normal
1-32	mild
33-58	moderate
59-132	severe

Table 4: UPDRS III Total

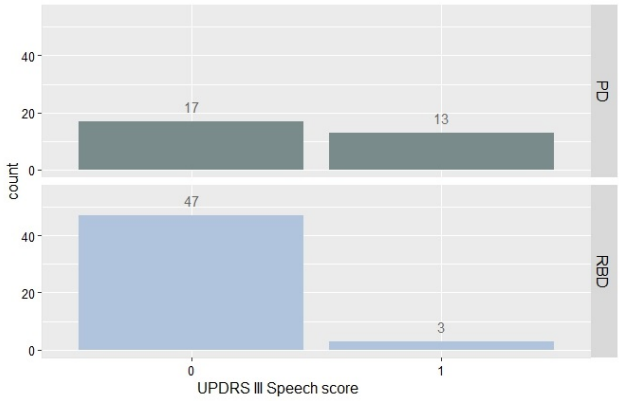


Within the third section of the UPDRS, there is only one item that concerns speech (*UPDRS III Speech*). As our project aims to analyze speech patterns in patients with Parkinson's disease (PD) and REM sleep behavior

disorder (RBD), it was important to assess their scores on this particular item.

0	normal
1	slight
2	mild
3	moderate
4	severe

Table 5: UPDRS III Speech



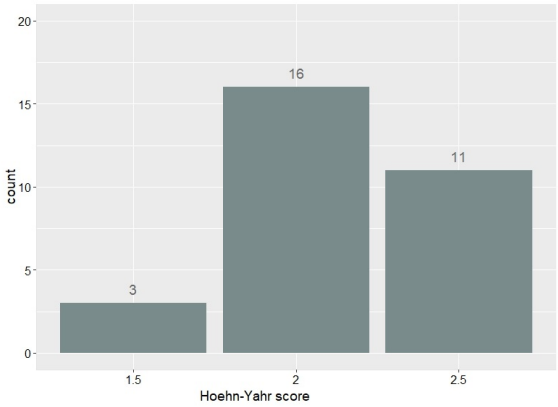
During the examination, clinicians observed slight speech alterations (score 1) in almost half of PD patients, while the other half were classified as 'normal' (score 0). On the other hand, no speech alterations were detected in almost all RBD patients.

4.3. Hoehn and Yahr scale

Another common way to assess the progression of Parkinson's disease is to adopt the Hoehn and Yahr scale. This scale was designed to be a descriptive staging scale to evaluate disability and impairment related to clinical disease progression. It assigns a value ranging from 0 to 5, where 0 means no signs of the disease, and 5 indicates a bedridden or wheelchair-bound patient.
In our dataset, only PD patients were evaluated using this scale. This is unlike the UPDRS scores, which are provided for both PD and RBD patients.

0	no signs of disease
1	unilateral involvement
1.5	unilateral and axial involvement
2	bilateral symptoms without balance impairment
2.5	mild bilateral disease with balance impairment
3	mild to moderate bilateral disease with postural instability
4	severe disability, but able to walk or stand unassisted
5	wheelchair-bound or bedridden unless assisted

Table 6: Hoehn and Yahr scale



The plot above shows that most PD patients have reached the stage of bilateral involvement, which implies the presence of symptoms on both sides of the body. Furthermore, some of these patients also experienced balance impairment, indicating that their motor control was already significantly affected. Out of all subjects affected by Parkinson's disease, only three patients have displayed symptoms that are limited to one side of their body (unilateral) or related to their neck and spine regions (axial).

5. Preprocessing

5.1. Non-speech-related variables

Before considering the speech-related features, we briefly analyzed the clinical variables and the disease progression scores in order to explain possible anomalies in the observations and get an overall understanding of the patients' scenario.

1. As a preliminary step, we analyzed the information relative to the medication. Given that only a small number of patients in the cohort were registered to be undergoing antidepressant therapy or taking benzodiazepines, we deemed that such variables were not informative for our purposes.
2. Then, we considered the motor ability evaluation variables. Since these observations obviously were not available for healthy patients, we performed several tests to assess whether there was a difference in the distribution between PD and RBD groups.
   Before performing such tests, in order to reduce the dimensionality without losing too much information, we merged variables related to the same motor symptoms by summing them (overall tremor, overall rigidity, ...), and then we applied permutational ANOVA. As expected, the testing procedure highlighted a clear distinction between Parkinson's and RBD patients for these variables.

5.2. Outliers Detection

Before describing the procedure that we used to identify possible outliers, it is right to say that the data collection process that generated our dataset is rigorous in the clinical sense, and if some outliers were present, the cause might be rather due to the low number of samples it contains. Indeed, the speech variables were extracted by an automatic pattern recognition system, as mentioned in the introduction, which is unlikely to produce anomalous registrations.

Since the ultimate goal of our project is to build a model to quantify the risk for an RBD patient to develop Parkinson's disease by means of speech patterns, we only looked for outliers in the speech-related variables for Control patients and Parkinson's patients.

We proceeded to set up a detection method by considering bivariate bagplots for each combination of 2 variables among the 24 speech-related features (12 features for each of the two tasks) and labeling as outliers those units that are outlying in at least one combination.
Because the dimensionality was high, the overall number of outliers spotted was 41 out of 80 units. Of course, this was not an acceptable result, so we decided to adopt a more rigid criterion and define as outliers those data points that are outlying in at least 25 combinations (about 10% of the total number of combinations). By applying this procedure, we detected 14 outliers.

After a more rigorous analysis of the outlying observations, we did not find evident incompatibilities with the regular units for what concerns clinical and demographic information, which we thought could be responsible for the difference. This fact supports the preface of this section: the anomalies found do not fit perfectly with the bulk of our dataset, but if we had more observations it would have been unlikely for those units to be labeled as outliers.

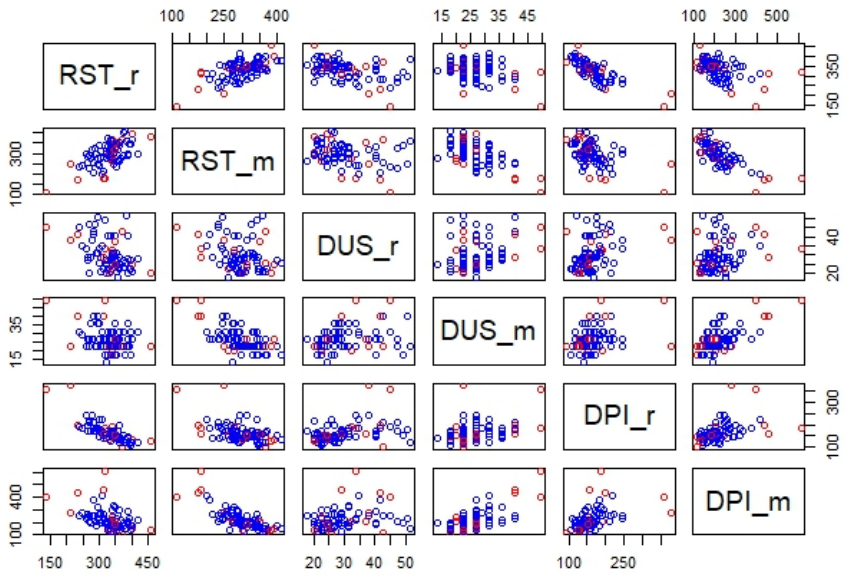


Figure 2: Outliers plot for the most relevant speech features. The outliers are marked in red and the regular observations in blue.

6. Feature selection

In order to achieve the final goal of accurately classifying RBD patients as either Parkinson's disease (PD) patients or healthy controls (HC), the first step in the feature selection process was to remove any variables that did not show a significant difference in distribution between the two categories (PD and HC). This was done to ensure that only the most informative features were used in the model, which would ultimately lead to better classification accuracy and reliability.

To assess whether we needed to use nonparametric methods or parametric ones were enough, we conducted two tests. Firstly, we tested whether the features follow a normal distribution (using the Shapiro-Wilk test). For those features that did follow a normal distribution, we moved on to test for homogeneity of variance in the two categories (using the F-test). We found that the following variables are normally distributed with homogeneous variance: AST_r, GVI_r, RLR_r, RST_m, GVI_m, DUF_m, RLR_m, RSR_m.

These were the first steps to determine which variables could be compared using a parametric t-test. For the variables identified above, we then applied the parametric t-test to evaluate whether the means of the two populations (the two categories, PD and HC) differ.
At significance level of $\alpha = 5\%$, we rejected the null hypothesis H_0 only for the variable RST_m (Rate of speech timing for the monologue task). In other words, only RST_m showed a significant difference in the mean of PD and HC categories.

For all the other variables, not suitable for the parametric t-test, we performed a permutation test using as test statistics the absolute value of the difference between the medians of the two groups, finding that RST_r, DPI_r, DUS_r, DPI_m are significant (at level $\alpha = 5\%$).

	p-value
AST_r	0.8708
GVI_r	0.6969
RLR_r	0.6489
RST_m	0.0026
GVI_m	0.5065
DUF_m	0.2822
RLR_m	0.2724
RSR_m	0.1051

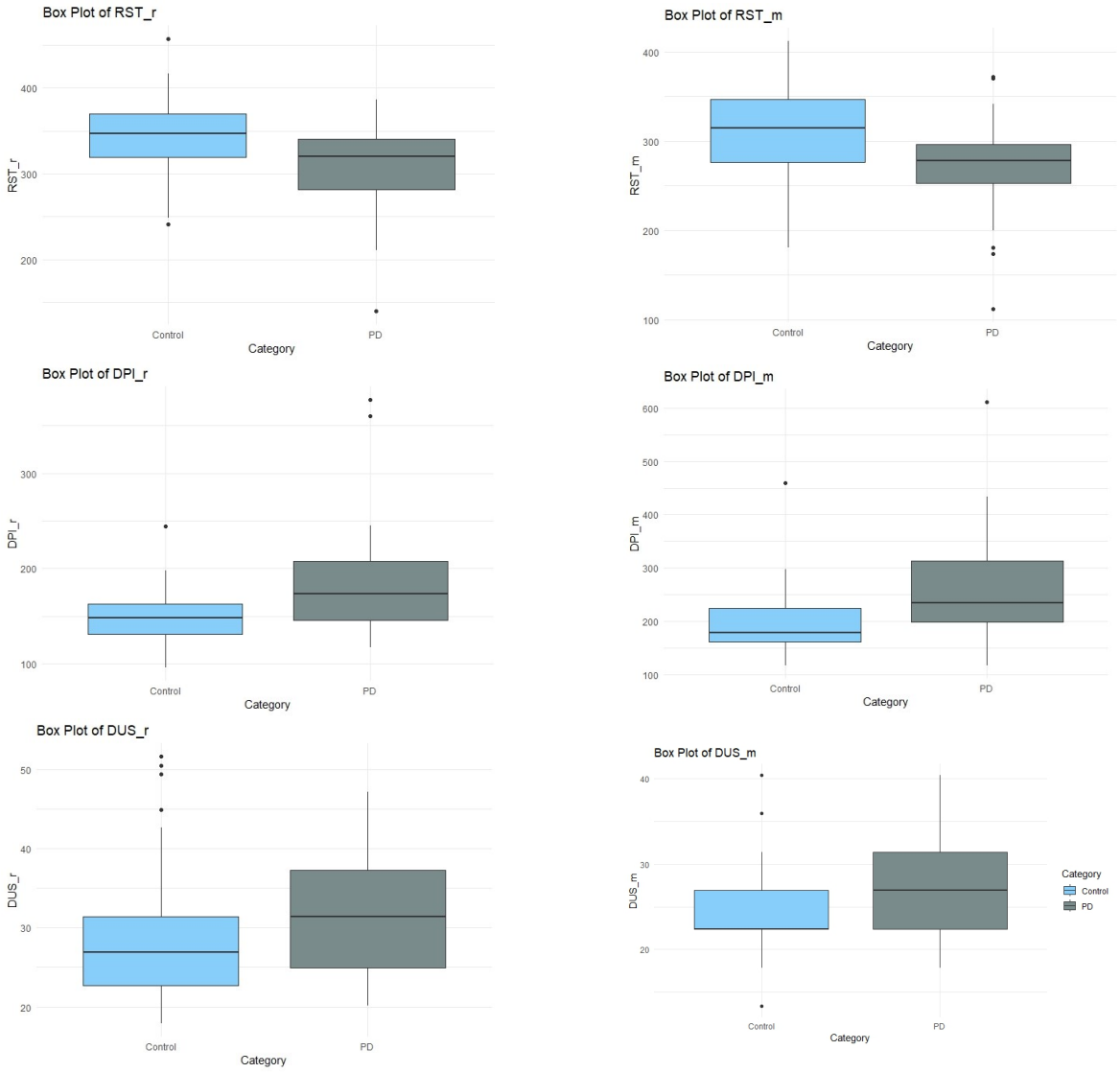
Table 7: parametric t-test

	p-value		p-value
EST_r	0.746	RST_r	0.005
DPI_r	0.001	DVI_r	0.065
DUS_r	0.020	DUF_r	0.622
PIR_r	0.961	RSR_r	0.544
LRE_r	0.952	EST_m	0.699
AST_m	0.245	DPI_m	0
DVI_m	0.204	DUS_m	0.162
PIR_m	0.586	LRE_m	0.074

Table 8: permutation test

Overall, this preliminary feature selection step reduced the number of speech features from 24 to 5. We used these five features, together with Age and Gender variables, to build models in the next steps.

In order to always keep in mind the influence of outliers, we repeated the same procedure with the version of the dataset without outliers. The variables selected in this case were the same as in the previous scenario, with the only difference that instead of DUS_r the result underlined a significant effect of DUS_m.



Another method that we adopted to extract the relevant speech features was to perform LASSO regularization on Generalized Linear Models with all the speech variables, setting as response the Parkinson's group belonging. The optimal value for the regularization parameter λ was obtained through a Leave-One-Out Cross-Validation procedure yielding as significant covariates only DPI_r and DPI_m.

In conclusion, we can state that different feature selection methods, both statistically driven and brute-force, highlighted that the most impactful speech markers of Parkinson's disease are related to the timing and articulation of the speech.

7. Regression models

The main goal of our project is to build a model to predict the risk for an RBD patient to develop Parkinson's disease using the speech markers identified in the previous steps. As the desired output of the model is the probability for a patient to be classified as affected by Parkinson's disease, we focused on the family of **Binary Logistic Regression** models. We tried several approaches that we will describe in the following subsections:

- Parametric models
- Semiparametric and nonparametric models
- Robust regression models

We defined the training set as the union of Control patients and Parkinson's patients. From here comes the binary label (1 for Parkinson's, 0 for Control) to be predicted by the model. The test set is given by the RBD patients, and to validate the model, we used the Leave-One-Out Cross-Validation technique. As the training set is composed of 66 units, the computational cost of the LOOCV score was not unbearable.
Given that the output of the model is a probability, we needed to choose a threshold to translate the risk (value in $[0, 1]$) into a tentative diagnosis ($0, 1$ label). We were flexible about this hyper-parameter and tried intermediate values between 0.5 and 0.7.
To assess the accuracy of the classification of the observations in the test set, we compared the results with the ones obtained by the UPDRS III* classifier with a confusion matrix.
The results on the test set varied according to different choices of the threshold. This probably happened also because of the class imbalance that tends to impose a bias on the Control category, which is represented by more samples.

A clarification has to be made before presenting the performance of our models in the form of confusion matrices. The UPDRS III* classification method evaluates the motor impairment of the patients. If this quantity is higher than the cut-off value of 3, the patient is marked as "at risk".
The goal of our models is to predict a risk factor that is not related to the motor impairment of the patients, so we are tackling the problem from a different perspective. Moreover, we are not training a hard-labeling classifier, but we are using a soft-labeling procedure.
This means that the comparison with the UPDRS III* classifier can only be an approximate measure of how well our models are grasping reality.

7.1. Parametric Logistic Regression

For this type of model, we adopted a brute-force approach. It means that as a preliminary step, we tried all the possible models coming from the selected features and kept only the best ones in terms of the LOOCV score. The best model reached a score of approximately 71%, which is not surprisingly high, but since the model is probabilistic, one needs to be flexible with the results.

This step was useful both to remove a priori models with poor performances and to identify which variables, among the ones selected previously with permutational inference, could be defined as early markers of the disease. According to this procedure, the most significant variables are *Gender*, Duration of unvoiced stops for the monologue task (DUS_m), and Duration of paused intervals for the monologue task (DPI_m).

After this initial step, we proceeded to consider from a statistical point of view only models built using these three significant variables. Specifically, we fitted a model with all 3 variables (*Gender*, DUS_m, DPI_m) and a reduced model with only *Gender* and DPI_m, and looked at the parametric significance of the covariates. Below are the mathematical formulations of the two models and the relative coefficients' tables.

$$\mathcal{M}_1 : Pr(Y = 1 \mid DUS_m, DPI_m, Gender_{male}) = \frac{e^{\beta_0+\beta_1 \cdot DUS_m+\beta_2 \cdot DPI_m+\beta_3 \cdot Gender_{male}}}{1 + e^{\beta_0+\beta_1 \cdot DUS_m+\beta_2 \cdot DPI_m+\beta_3 \cdot Gender_{male}}}$$

	Estimate	Std. Error	z value	$\mathbb{P}(>\|z\|)$
Intercept	-4.46773707	1.598333964	-2.7952463	0.005186019
DUS_m	0.04912476	0.055641980	0.8828722	0.377305325
DPI_m	0.01602978	0.006451571	2.4846323	0.012968533
Gender = Male	-1.28458074	0.717531236	-1.7902785	0.073409152

Table 9: Full model

$$\mathcal{M}_2 : Pr(Y = 1 \mid DPI_m, Gender_{male}) = \frac{e^{\beta_0+\beta_1 \cdot DPI_m+\beta_2 \cdot Gender_{male}}}{1 + e^{\beta_0+\beta_1 \cdot DPI_m+\beta_2 \cdot Gender_{male}}}$$

	Estimate	Std. Error	z value	$\mathbb{P}(>\|z\|)$
Intercept	-3.63391837	1.257730950	-2.889265	0.003861432
DPI_m	0.01835659	0.006039795	3.039273	0.002371500
Gender = Male	-1.33746409	0.709909111	-1.883993	0.059565866

Table 10: Reduced model

We observe that in the full model, the covariate DUS_m is not significant. For this reason, a reduced model was fitted, considering only *Gender* and DPI_m. We compared the full and the reduced model using an ANOVA test, and it confirmed that a reduced model is preferred.

By fixing the threshold at 0.5, the reduced model provides the following confusion matrix on the test set:

	Predicted HC	Predicted PD
UPDRSIII HC	17	10
UPDRSIII PD	14	9

Table 11: Confusion Matrix ($\mathcal{M}_2$) — threshold 0.5

By changing the threshold to 0.6, we get an increase in the performance on the test set, even if the LOOCV score remains the same on the training set.

	Predicted HC	Predicted PD
UPDRSIII HC	22	5
UPDRSIII PD	16	7

Table 12: Confusion Matrix ($\mathcal{M}_2$) — threshold 0.6

Below we report the regression lines of $\mathcal{M}_2$ for male and female patients.
One can clearly see that the risk is increased for female patients.

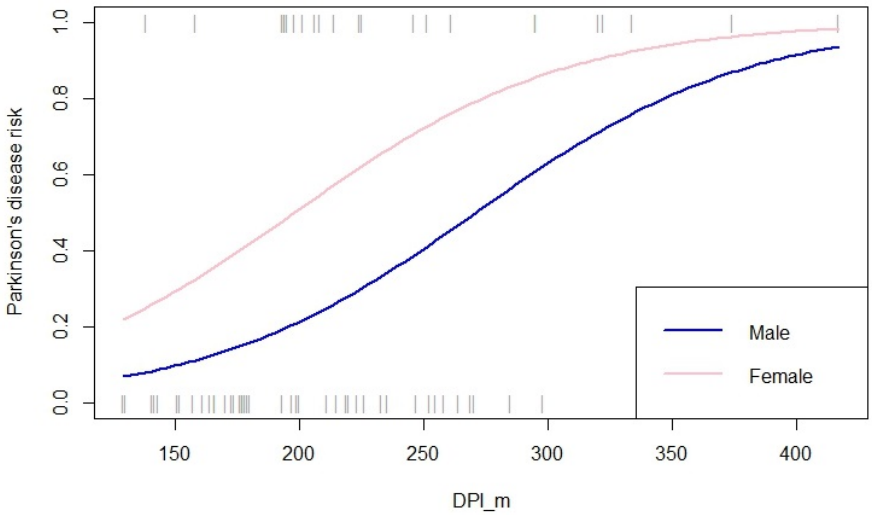


Figure 3: $\mathcal{M}_2$ regression plot

We can conclude this paragraph by stating that the parametric models are probably not the ideal choice, given that the best model contains only one speech-related significant covariate. However, they reflect the general trend that we deepen in the next sections.

7.2. Semiparametric and Nonparametric Logistic Regression

Seeing the non-optimal results obtained by the classical parametric models, we decided to broaden the horizon by including semiparametric and nonparametric effects in the models.
Since our training set is not large, we could not fit high-dimensional models, as these would have surely suffered from overfitting issues. Hence, we considered possible combinations of smoothing functions of the variables selected previously in the feature selection section, building different kinds of Generalized Additive Models (GAM). The best models are described below.

$$\mathcal{M}_3 : Pr(Y = 1 \mid Gender_{male}, RST_r, DPI_m) = \frac{e^{\beta_0+\beta_1 \cdot Gender_{male}+\beta_2 \cdot RST_r+f(DPI_m)}}{1+e^{\beta_0+\beta_1 \cdot Gender_{male}+\beta_2 \cdot RST_r+f(DPI_m)}}$$

This model was fitted on the training set *without the outliers*. The smoothing function f is given by cubic regression splines. All the terms of the model are statistically significant at level 10%.

	Predicted HC	Predicted PD
UPDRSIII HC	22	5
UPDRSIII PD	13	10

Table 13: Confusion Matrix ($\mathcal{M}_3$) — threshold 0.5

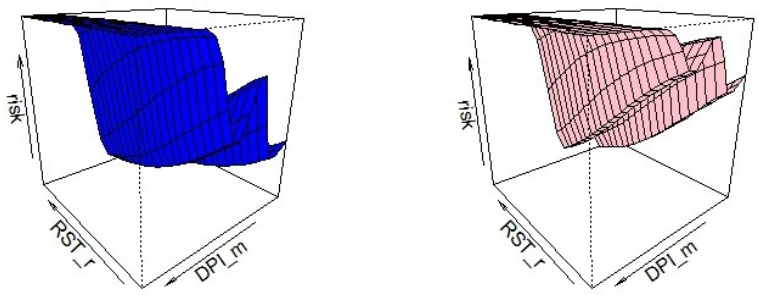


Figure 4: $\mathcal{M}_3$ logistic surfaces for male and female patients

$$\mathcal{M}_4 : Pr(Y = 1 \mid Gender_{male}, RST_m, DPI_r) = \frac{e^{\beta_0+\beta_1\cdot Gender_{male}+\beta_2\cdot RST_m+f(DPI_r)}}{1+e^{\beta_0+\beta_1\cdot Gender_{male}+\beta_2\cdot RST_m+f(DPI_r)}}$$

This model was fitted on the *full training set*. The smoothing function f is given by thin-plate regression splines. All the terms of the model but the intercept are statistically significant at level 5%. Even if the intercept is not significant, we can't remove it from the model, as this would be equivalent to assuming that $\mathbb{P}(Y = 1) = 0.5$ when all the predictors are zero. In our case, we can't make such an assumption, so we keep the intercept.

	Predicted HC	Predicted PD
UPDRSIII HC	18	9
UPDRSIII PD	11	12

Table 14: Confusion Matrix ($\mathcal{M}_4$) — threshold 0.5

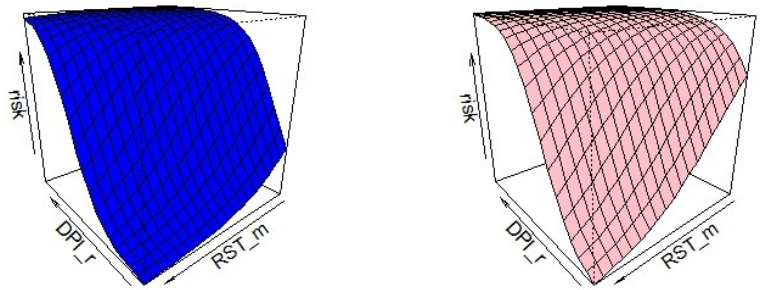


Figure 5: $\mathcal{M}_4$ logistic surfaces for male and female patients

The selection of the covariates for these two models has been made considering statistical significance, cross-validation scores, correlation of the linear terms, and concurvity of the smoothing terms.
Concurvity occurs when some smooth term in a model could be approximated by one or more of the other smooth terms in the model. High concurvity among covariates might make estimates unstable and consequently cause problems in the interpretation of the model.

To predict the risk of an RBD patient developing Parkinson's disease, using the two models described above, we have employed both the 'classical' `predict` method as well as the full conformal prediction approach. Both prediction procedures yielded the same results in terms of point-wise risk predictions and confusion matrices. Considering the heavy computational cost associated with full conformal prediction, we opted for the `predict` method instead.

7.3. Robust Regression

To understand if our outlier detection algorithm was effective and to somehow unify the results of our two best models, which were fitted respectively with and without considering the outliers, we performed a robust analysis procedure.

To identify the outliers, we first computed the robust estimators of location and scatter of the speech features for the dataset composed of Parkinson's and Control patients by applying the Minimum Covariance Determinant method (and setting `alpha=0.75, nsamp="best"`). Then, the outliers were identified based on their large squared Mahalanobis distance from the robust fit.
Overall, using the MCD approach, we identified 14 outliers, which is consistent with our previous findings, as the number of spotted outliers is the same. However, there are a few differences in the observations labeled as outliers, as shown in the table below.

MCD	3	4	6	8	11	18	23	28	33	34	41	50	52	—	72	—	—
outlier detection algorithm	3	4	6	8	11	—	23	28	—	—	41	50	52	59	72	74	75

Table 15: Row indexes of the outlying observations

As the next step, we proceeded by fitting Generalized Linear Models in a robust fashion using the R package `robustbase`. The best model is the following:

$$\mathcal{M}_5 : Pr(Y = 1 \mid Gender_{male}, DPI_r, DPI_m) = \frac{e^{\beta_0+\beta_1 \cdot Gender_{male}+\beta_2 \cdot DPI_r+\beta_3 \cdot DPI_m}}{1 + e^{\beta_0+\beta_1 \cdot Gender_{male}+\beta_2 \cdot DPI_r+\beta_3 \cdot DPI_m}}$$

All the terms of the model are significant at level 10%, and the overall performance is very close to one of the models seen previously. For the sake of completeness, we report below the plots of the logistic surfaces and decision boundaries, separately for male and female patients.

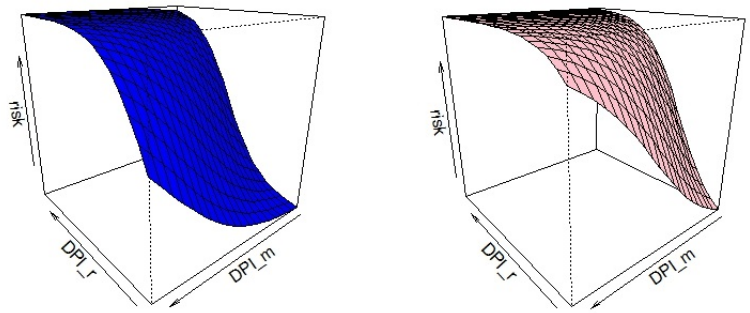


Figure 6: $\mathcal{M}_5$ logistic surfaces for male and female patients

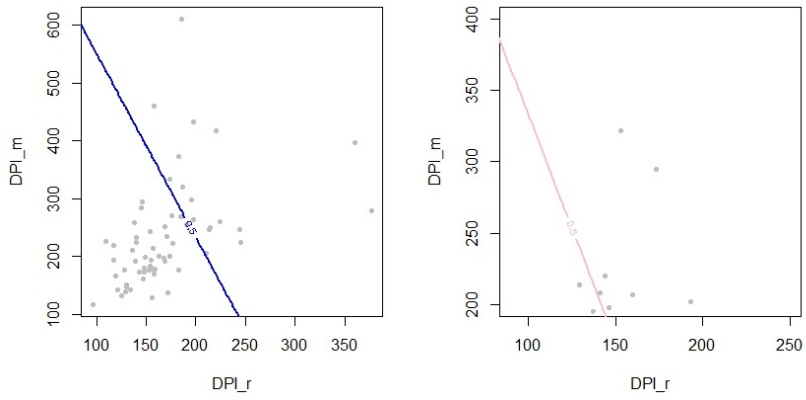


Figure 7: $\mathcal{M}_5$ decision boundaries for male and female patients

7.4. Speech-related markers

By looking at all the regression models we developed, regardless of the statistical method adopted to build them, it is clear that some features are more relevant than others in assessing the probability of a Parkinson's diagnosis.

The results can be summed up as follows:

- High duration of paused intervals when producing monologues or reading a written passage is linked with higher risk. Indeed, complex speech impairment can cause difficulties in initiating speech, which induces the extension of pauses.
- Lower values of rate of speech timing are associated with higher risk. Since speech rate impairment is related to deficits in all dimensions of speech, this measure is useful because it merges voiced intervals, unvoiced intervals, and pauses.

- Both the reading and monologue tasks are relevant, but we expect high-risk patients to struggle more in producing speech without any landmarks rather than reading a pre-existing passage.

8. An alternative approach: cluster analysis

To tackle the problem from a different perspective, we performed a cluster analysis on the RBD patients, considering the variables obtained in the preprocessing step. The goal of this part is to identify a specific subgroup that shows similarities with PD patients.

8.1. Hierarchical clustering

The first algorithm used was hierarchical clustering. We computed the similarity matrix among patients with Euclidean distance, and the dendrogram was constructed using Ward linkage. Below is the resulting dendrogram.

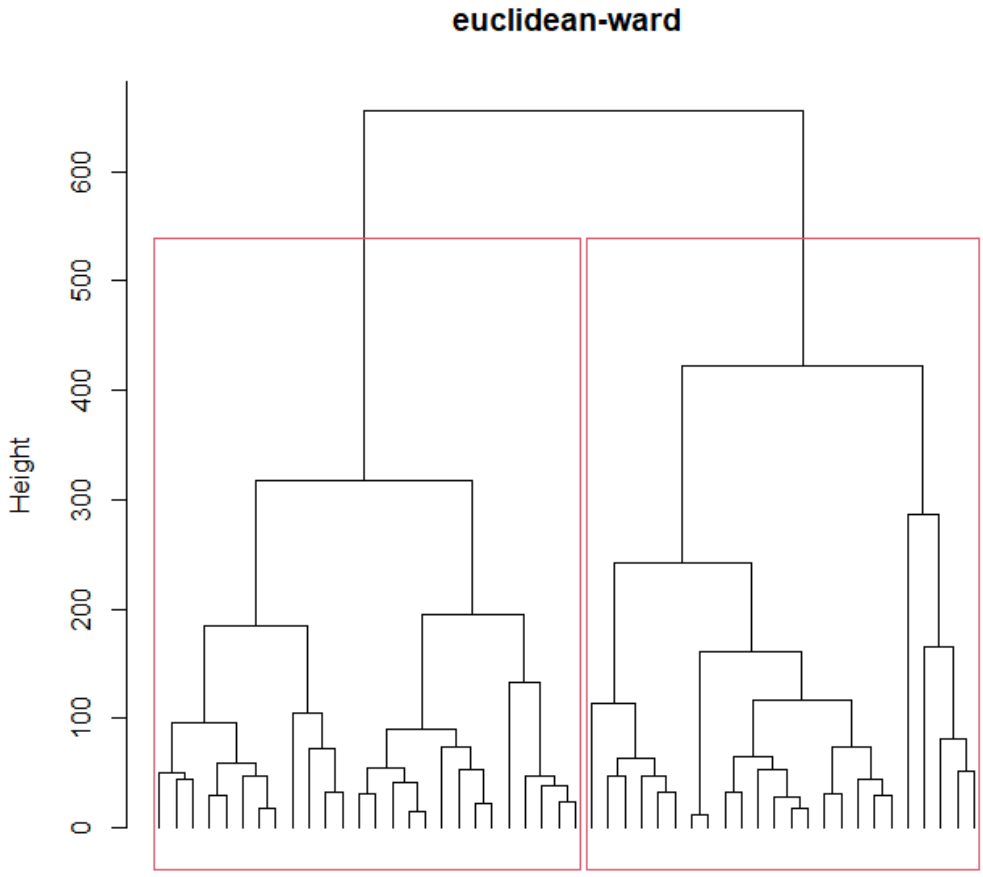


Figure 8: Dendrogram

After analyzing the dendrogram, we have concluded that it would be appropriate to create two clusters. With these two clusters, we wanted to identify two distinct patterns in RBD patients: those closer to the PD group and those more similar to the Control group.
To validate the quality and reliability of the clustering, we performed a permutational MANOVA. It allowed us to evaluate whether the observed differences in the two clusters were significant and not due to random variation. Here are the hypotheses of the test:

H_0: cluster_1 and cluster_2 don't show differences;
H_1: there is a difference between them.

The p-value is equal to 0, so we assumed that the clustering was well done.

To determine which cluster is closer to the PD group, we performed two permutational MANOVAs, one for each of the two clusters. For each comparison, we considered the following test:

H_0: cluster$_i$ is similar to PD patients;
H_1: cluster$_i$ is not similar to PD patients;
with $i = 1, 2$

When including the outliers in the data, we obtained a p-value of 0.9% for cluster 1 and 2.5% for cluster 2. So, at a significance level of $\alpha = 1\%$, the former shows more similarity with Control patients than the latter. Below is the relative confusion matrix, obtained by labeling cluster 1 as 'HC' and cluster 2 as 'PD'.

	Predicted HC	Predicted PD
UPDRSIII HC	17	10
UPDRSIII PD	9	14

Table 16: Hierarchical clustering with outliers

Instead, when focusing on the data without the outliers, we obtained a p-value of 3.5% for cluster 1 and 3.2% for cluster 2. At the same level of significance ($\alpha = 1\%$), it is not clear which of the two clusters can be considered as closer to Parkinson's patients. Since the former is associated with a slightly higher p-value, we considered it as PD-like. Here the resulting confusion matrix is the following:

	Predicted HC	Predicted PD
UPDRSIII HC	10	17
UPDRSIII PD	9	14

Table 17: Hierarchical clustering without outliers

We are aware that the last result is not reliable.

8.2. K-means

Keeping the same strategy as before, we tried another clustering algorithm: K-means. The permutational MANOVA test to assess the quality of the clustering (two clusters) returned a p-value equal to 0. Hence, also in this case, we considered the clustering to be well done.

Next, we performed the two tests to identify PD patterns in the two clusters. When including the outliers in the data, we obtained a p-value of 0.5% for cluster 1 and 2.5% for cluster 2.

	Predicted HC	Predicted PD
UPDRSIII HC	17	10
UPDRSIII PD	9	14

Table 18: K-means clustering with outliers

We observe that the confusion matrix is exactly the same as the one obtained using hierarchical clustering (with outliers).

The problems came out when we considered the data without the outliers: the quality of clustering was still deemed as good, as the p-value of the relative test was 0. However, the pattern identification tests returned the

p-values of 3.8% for cluster 1 and 4.2% for cluster 2. Similarly to what happened with hierarchical clustering, we were not able to conclude which one is closer to PD patients.
Using the first cluster as Control-like, we obtained the following confusion matrix:

	Predicted HC	Predicted PD
UPDRSIII HC	8	19
UPDRSIII PD	13	10

Table 19: K-means clustering without outliers

We can notice that the number of misclassified patients is significantly higher with respect to what we have obtained before.

8.3. Limits of this approach

The previous results show an important limit of this approach: we can notice that as we remove the outliers, the reliability of the results decreases significantly. A possible motivation is that, given the small number of observations and the fact that we were using an unsupervised algorithm, the outliers were fundamental to identifying more precise patterns in the two clusters.
On the other hand, the regression approach has shown good results with and without them: this can be noticed by looking at the results of our best regression model and the outcome of the robust regression.
This observation confirms how the techniques seen during the course are very powerful and lead to consistent results.

9. Conclusions

The goal of this project was to identify new markers that could allow us to assess the risk of an RBD patient developing Parkinson's disease. The Rapid Eye Movement Sleep Behavior Disorder is known to be a possible preamble for that condition, so anticipating the diagnosis could be crucial to be able to tackle it optimally.

Looking at the results of our analysis we can consider the goal accomplished. Indeed, we developed a tentative alternative diagnostic method that can be useful since it considers objective parameters that can be collected automatically. We believe that the statistical procedures we adopted are solid. Nevertheless, we are aware of possible flaws as the unbalance of the dataset in the gender variable or the low number of observations that could have influenced the final prediction of the model. However, these problems can be solved by repeating the same procedure on a richer and more balanced dataset, possibly yielding different models.

This analysis proves that the research has no boundaries: with the techniques available nowadays, we can contribute significantly to making important steps toward the new frontiers of medicine. These will allow clinicians to make more precise diagnoses and find solutions tailored to the specific needs of each patient.

A potential direction to further advance this research is to consider speech measurements over a certain period of time. By applying survival analysis techniques and functional data analysis to such measurements, one can investigate the time it takes an RBD patient to develop Parkinson's disease or the progression of the RBD and PD conditions over time. This can serve as an excellent starting point for further exploration of this topic and can lead to new insights and discoveries in this field.

References

[1] Hlavnička J., Čmejla R., Tykalová T., Šonka K., Růžička E., and Rusz J. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports*, 2017. URL: https://doi.org/10.1038/s41598-017-00047-5.

[2] Van Patten R., Mahmood Z., Pickell D., Maye J., Roesch S., Twamley E., Filoteo V., and Schiehser D. REM Sleep Behavior Disorder in Parkinson's Disease: Change in Cognitive, Psychiatric, and Functional Outcomes from Baseline to 16–47-Month Follow-Up. *Archives of Clinical Neuropsychology*, 2022. URL: https://doi.org/10.1093/arclin/acab037.

[3] Postuma R.B., Lang A.E., Gagnon J.F., Pelletier A., and Montplaisir J.Y. How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder. *Brain: a journal of neurology*, 2012. URL: https://doi.org/10.1093/brain/aws093.