Predicting Parkinson's disease trajectory using clinical and functional MRI features: a reproduction and replication study

Elodie Germani, Nikhil Baghwat, Mathieu Dugré, Rémi Gau, Albert Montillo, Kevin Nguyen, Andrzej Sokolowski, Madeleine Sharp, Jean-Baptiste Poline, Tristan Glatard

Predicting Parkinson's disease trajectory using clinical and functional MRI features: a reproduction and replication study

Elodie Germani[1,*], Nikhil Baghwat[2], Mathieu Dugré[3], Rémi Gau[2], Albert Montillo[4], Kevin Nguyen[4], Andrzej Sokołowski[3], Madeleine Sharp[2], Jean-Baptiste Poline[2,+], Tristan Glatard[3,+]

1 Univ Rennes, Inria, CNRS, Inserm, France
2 Department of Neurology and Neurosurgery, McGill University, Montreal, Canada
3 Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
4 Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, USA

+ Equal contributions

*** elodie.germani@irisa.fr**

Abstract

Parkinson's disease (PD) is a common neurodegenerative disorder with a poorly understood physiopathology. In clinical practice, challenges are encountered in the diagnosis of early stages and in the prediction of the disease progression due to the absence of established biomarkers. Several biomarkers obtained using neuroimaging techniques such as functional Magnetic Resonance Imaging (fMRI) have been studied recently. However, the reliability and generalizability of neuroimaging-based measurements are susceptible to several different sources of variability, including those introduced by different analysis methods or population sampling. In this context, an evaluation of the robustness of such biomarkers is essential. This study is part of a larger project investigating the replicability of potential neuroimaging biomarkers of PD. Here, we attempt to reproduce (same data, same method) and replicate (different data or method) the models described in [1] to predict individual's PD current state and progression using demographic, clinical and neuroimaging features (fALFF and ReHo extracted from resting-state fMRI). We used the Parkinson's Progression Markers Initiative dataset (PPMI, ppmi-info.org), as in [1] and tried to reproduce the original cohort, imaging features and machine learning models as closely as possible using the information available in the paper and the code. We also investigated methodological variations in cohort selection, feature extraction pipelines and sets of input features. Using the reproduction workflow, we managed to obtain better than chance performance for all our models ($R2 > 0$), but this performance remained very different from the ones reported in the original study. The challenges encountered while reproducing and replicating the original work are likely explained by the complexity of neuroimaging studies, in particular in clinical settings. We provide recommendations to facilitate the reproducibility of such studies in the future, for instance with the use of version control tools, standardization of pipelines and publication of analysis code and derived data.

Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder with more than 10 million people affected in the world. Disease manifestations are heterogeneous and their evolution varies between patients, dividing them in different subtypes and stages [2]. Identification of these stages or subtypes is essential for clinical trials as well as for clinical practice to track the disease progression, however, there is currently no established biomarker of disease severity or progression [3, 4].

Neuroimaging techniques are able to capture rich and descriptive information about brain structure and functional architecture non-invasively. In conjunction with computational algorithms based on pattern recognition and machine learning, neuroimaging measures began to emerge as candidate PD biomarkers in the past few years. Among other imaging modalities, functional Magnetic Resonance Imaging (fMRI), which estimates the blood oxygenation level-dependent (BOLD) effect to represent neural activity, showed a high potential in identifying specific biomarkers related to PD and its progression [5]. While disease phenotypes are heterogeneous, neuronal dysfunction patterns were shown to be highly replicable between patients [6].

Resting-state fMRI (rs-fMRI) features are particularly promising. Region-wise measurements such as regional homogeneity (ReHo) and Amplitude of Low Frequency Fluctuations (ALFF) were used in several studies to predict PD trajectory or motor subtypes [7, 8, 9, 1, 10, 11]. ReHo quantifies the connectivity between a voxel and its nearest neighboring voxels and was shown to be affected by neurodegenerative diseases [12]. ALFF and its normalized form, fractional ALFF (fALFF), measure the power of the low frequency signals at rest, which mostly consists in spontaneous neuronal activity [13].

However, despite their potential, neuroimaging measures are sensitive to multiple sources of variability that impact their reliability and may explain why the derived biomarkers are not well established in clinical and research practice. In particular, neuroimaging analyses require specific methodological choices at various computational steps, related to the software tools, the method, and the parameters to use. These choices, also known as "researchers' degrees of freedom" [14], might have a large impact on the results of an experiment and sometimes lead to a lack of agreement when analyzing the same neuroimaging dataset with different analysis pipelines [15, 16]. For instance, in task-based fMRI, 70 research teams were asked to analyze the same fMRI dataset using their usual analysis pipeline and results were substantially variable across teams [16].

Furthermore, neuroimaging results have been shown to be impacted by differences in hardware architectures or software package versions [17, 18]. This suggests that a single pipeline evaluation is not sufficient to obtain robust results, though the reliability of results may be increased when studying their distributions across perturbations.

There are also concerns about the reproducibility of machine learning studies. Indeed, in a recent study [19], researchers attempted to reproduce several machine learning experiments, revealing multiple issues which could lead to the non-reproducibility of findings. These issues can be split in three categories [20]: data leakage, computational reproducibility, and choice of evaluation metrics. In particular, [21] performed a review of CNN-based classification of Alzheimer's subtypes and found a potential data leakage in half of the 32 surveyed studies due to a wrong data split at the subject-level, a data split after data augmentation or dimension reduction, transfer learning with models pre-trained on parts of the test set or the absence of an independent test set. These data leakage might cause an over-optimistic performance assessment of models and thus, a lack of reproducibility and replicability of the findings. Evaluation procedures can also cause the non-reproducibility of findings, due to unsuitable metric choices when using unbalanced datasets for instance or questionable cross-validation procedures, in particular with low sample sizes. Random choices in a training procedure, for instance initial weights or hyper-parameters random selection, which all impact computational

reproducibility, might also lead to uncontrolled fluctuations in results when using different random initialization states.

Conflicting terminologies exist for the terms reproducibility and replicability [22]. Here, we define reproducibility as attempts made with the same methods and materials. Replicability, on the other hand, is tested with different but comparable materials or methods.

Replicability experiments have shown different degrees of variability between findings obtained with different analytic conditions. These studies are usually done using healthy populations and in general research practice (as opposed to clinical research), as in [16]. For clinically-oriented research, however, the topic remains understudied. Such studies requires a specific attention as they are useful to develop new biomarkers that can influence treatment development and clinical trial applications. These studies also often target specific populations of patients with unique characteristics, in particular for PD for which inter-individual variability is high [23]. Such studies often use small sample sizes, which has been shown to lead to a lower reproducibility of findings [24, 25]. Reproducibility and replicability of studies in clinical settings is of higher importance to improve the trustworthiness of new biomarkers and to facilitate their development.

In this paper, we evaluate the reproducibility and replicability of the study in [1], a clinically-oriented research on a PD population. The study in [1] is of particular interest as it uses the Parkinson's Progression Markers Initiative (PPMI) dataset [26], a large open access dataset to study Parkinson's disease. Moreover, it investigates the clinically relevant problem of trying to predict an individual's current and future disease severity over up to 4 years and it uses two different rs-fMRI-derived biomarkers: ReHo and fALFF. In [1], the authors trained several machine learning models using regional measurements of ReHo or fALFF along with clinical and demographic features to predict Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) total score at acquisition time and up to 4 years after. They selected n=82 PD patients by searching for those with rs-fMRI and MDS-UPDRS score at the same visit from the PPMI database and preprocessed the functional images to extract whole-brain maps of fALFF and ReHo. They compared three atlases, splitting the brains in different numbers of regions to extract mean region-wise features which are fed to the machine learning models. They achieved better than chance performance for prediction at each time point with both fALFF and ReHo, e.g. r-squared of 0.304 and 0.242 for prediction of current severity with ReHo and fALFF respectively. Finally, the authors discussed the most important brain regions for prediction and evaluated the performance of their models on an external dataset.

Different criteria could be used to conclude on success of the reproduction and replication of this study: 1) if the models trained on fALFF and ReHo at each time points showed better than chance performance in terms of r-squared (R2 >0 and R2 >chance-model R2) when tested on the PPMI dataset using the evaluation procedure proposed in [1] and 2) if these models showed similar performance (R2 greater than 0 and absolute difference between original and reproduction R2 less than 0.2) to those proposed in the original study. Our main interests were to assess the difficulties and challenges of reproducing fMRI research experiments, but also to evaluate the impact of different analytical choices on the results of these experiments. In this paper, we explore how these choices affect different parts of the analysis:

- Cohort selection and sample size,
- fMRI pre-processing pipeline,
- fMRI feature quantification,
- Choice of input features for machine learning models,

- Machine learning models choice and results reporting.

We also discuss the difficulties encountered to reproduce neuroimaging studies, in particular in clinical research settings, and we provide some recommendations on how to facilitate the reproducibility of such studies in the future.

Materials and Methods

Our study consisted of two steps: a first replication attempt without contacting the authors, using only publicly-shared resources available with the original paper, and a second replication attempt after contacting the authors, to obtain more accurate information on the original study. This two-step reproduction was meant to assess the challenges of reproducing a study using only publicly available materials and to evaluate the contribution of data and code sharing platforms to results reproducibility.

Dataset

As in the original study, we used data available from the Parkinson's Progression Markers Initiative (PPMI) dataset [26], a robust open-access database providing a large variety of clinical, imaging data and biologic samples to identify biomarkers of PD progression. The PPMI study was conducted in accordance with the Declaration of Helsinki and the Good Clinical Practice (GCP) guidelines after approval of the local ethics committees of the participating sites. We signed the Data User Agreement and submitted an online application to access the data. More information about study design, participant recruitment and assessment methods can be found in [26].

Summary of experiments

Reproducing an analysis can be challenging due to (1) the lack of specific information on analysis pipelines, software versions, or specific parameter values, (2) the presence of confusing terms in the available information, (3) the evolution of the software and data materials used in the original study. Our reproduction study consisted of 5 global steps: cohort selection, image pre-processing, imaging features computation, choice of input features and model choice and reporting. We used the information available in the original paper and for some parts of the analysis, we also had access to the code shared by the authors on GitHub (e.g. for feature computation and machine learning models). At each step, we still had to make informed guesses due to the 3 types of challenges stated above, which resulted in a high number of possible workflows. To evaluate the effect of each variation at each step, we defined a *default workflow* to which the variation was compared to. At each step, if a variation of the workflow was tested, the other steps were implemented as in the default one. This default workflow was the most likely according to the code shared along with the paper. Figure 1 summarizes the different variations tested and the *default workflow*. In the remainder, we compare the variations in the workflows based on statistical differences on participants features for cohort selection and on the models performance for other variations using several statistic tests implemented using scipy v1.11.2 (RRID:SCR_008058) [27].

Cohort selection

The cohort reported in [1] consisted of 82 PD participants with rs-fMRI and MDS-UPDRS scores obtained during the same visit. MDS-UPDRS Part III (motor examination) was conducted when patients were under the effect of PD medication. Of these 82 participants,

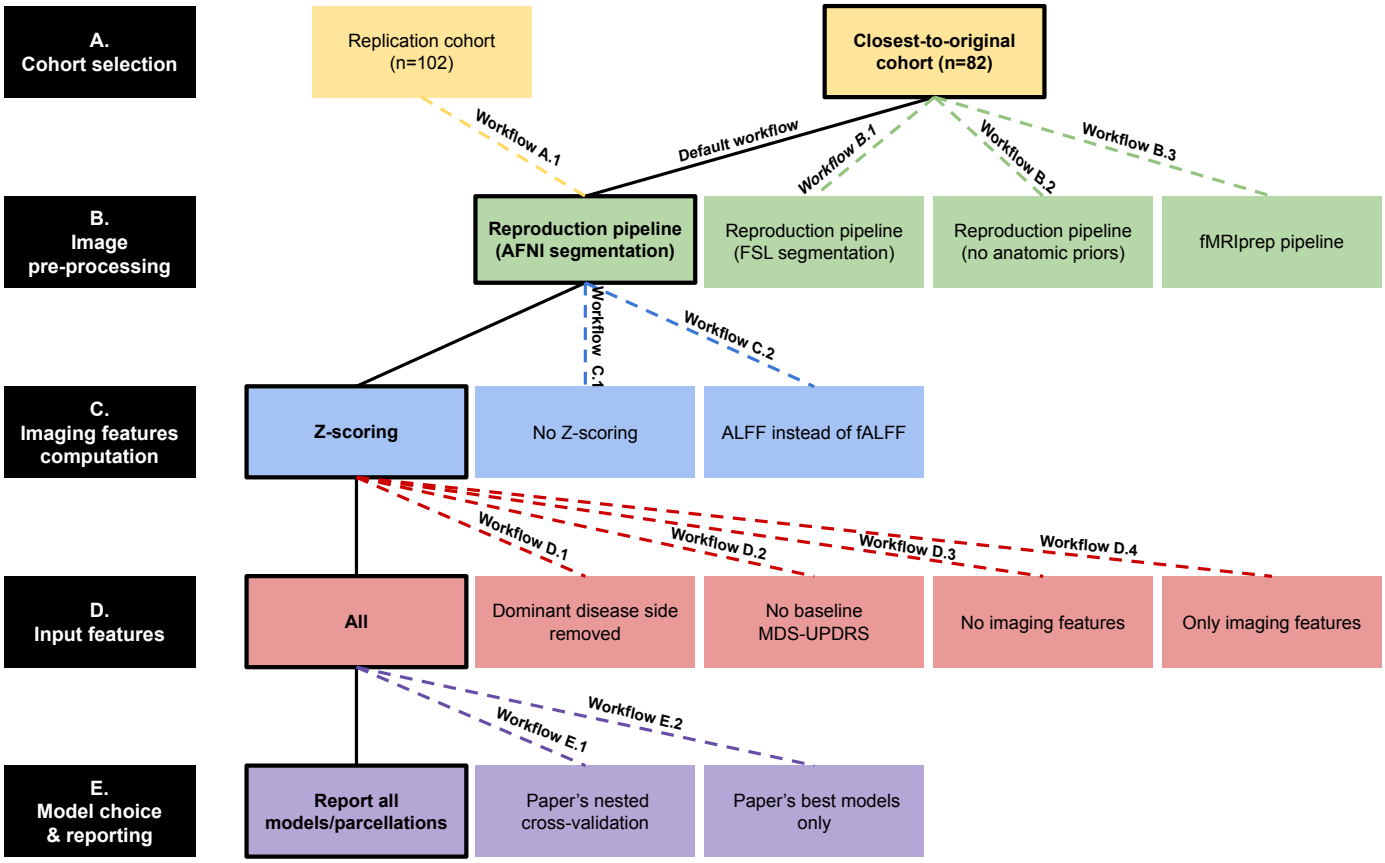

Figure 1. Summary of the different workflows implemented to reproduce the results of [1] and explore their robustness to different analytic conditions. Bold and bordered cells represent the implementation of the default workflow at each step, this whole workflow is labeled *Default workflow* and is represented using a plain bold line. The different variation workflows are represented in dashed lines: all steps different from the variation follow the default workflow and each workflow corresponds to one variation from the default one.

Variations of cohort selection (A):
- *Workflow A.1* - default workflow with replication cohort.

Variations of pre-processing pipeline (B):
- *Workflow B.1* - default workflow with FSL segmentation,
- *Workflow B.2* - default workflow without structural priors,
- *Workflow B.3* - fMRIprep pipeline.

Variations of feature computation (C):
- *Workflow C.1* - default workflow with no Z-scoring,
- *Workflow C.2* - default workflow with ALFF.

Variations of input features (D):
- *Workflow D.1* - default workflow with no dominant disease side,
- *Workflow D.2* - default workflow with no Baseline MDS-UPDRS,
- *Workflow D.3* - default workflow with no imaging features,
- *Workflow D.4* - default workflow with only imaging features.

Variations in model choice and reporting (E):
- *Workflow E.1* - default workflow with paper's nested cross-validation,
- *Workflow E.2* - default workflow with only paper's best model reporting.

53 participants also had MDS-UPDRS scores available at Year 1 after imaging, 45 at Year 2, and 33 at Year 4.

Replication cohort

We first attempted to reproduce the cohort of [1] using only the information available in the code shared on GitHub and the paper. Based on this information, we filtered the PPMI database using 4 criteria:

- Participants belong to the "Parkinson's disease" cohort, as defined in PPMI.
- Participants have an fMRI acquisition and a MDS-UPDRS score, with MDS-UPDRS Part III conducted ON-medication ("PAG_NAME" different from "NUPDRS3" in the PPMI score file) computed at the same visit (same visit code in PPMI database). Thus, only participants with valid values for MDS-UPDRS Part III score were included in the cohort.
- Participants and visits were also filtered depending on the type of fMRI acquisition. We queried the database with the exact same information as in the S1 Table of the original paper (field strength = 3T, scanner manufacturer = Siemens, pulse sequence = 2D EPI, TR = 2400ms, TE = 25ms).
- We also filtered the database to keep only participants for which the visit date and archive date of the image was set before January 1st, 2020 (more than a year before the original study publication) since we had no information about the date the authors accessed the database.

This query involved both fMRI metadata obtained using a utility functions from the Python packages livingpark-utils v0.9.3 and ppmi_downloader v0.7.4 and the MDS-UPDRS-III file from the PPMI database.

We queried the PPMI database on August 21st, 2023 and we included the participants selected using these filters in the Baseline time point of our replication cohort. To find the participants who also had a score available at Year 1, Year 2, or Year 4 follow-up, we looked for the visit date associated with the MDS-UPDRS score at Baseline and searched for participants that also had a score at 365 days (1 year) +/- 60 days (2 months), 2 $\times$ 365 days (2 years) +/- 60 days (2 months) and 4 $\times$ 365 days (4 years) +/- 60 days (2 months). This method was also used by the original authors to search for their cohort at Year 1, Year 2, and Year 4 follow-up.

Closest-to-original cohort

After contacting the authors, the exact participant and visit list used at Baseline was provided to us. We queried the PPMI database using this list and compared with our replication cohort.

The 82 participants of the original Baseline cohort were all included in our replication cohort. For 4 of them, the visit used in our replication cohort was different from the one used in the original cohort. For two participants, we used an earlier visit than the authors: V06 (2 years) instead of V10 (4 years) and BL (baseline) instead of V04 (1 year). For the last two participants that had different visits selected in the replication cohort, images of the visits used by the original authors were not available in the PPMI database when we queried it. We assumed this issue resulted from the update of the PPMI database in September 2021.

The 82 participants of the original cohort that were also included in our replication cohort were used to build a "closest-to-original" cohort to compare with our original cohort. The authors also provided the participant identifiers included at Year 1, Year 2

and Year 4, but we did not have the exact visit used at these time points. Thus, for each time point, we searched for the participants involved in our replication cohort for this time point that were in the list provided by the authors. Several participants from the list provided by the authors were not found in our cohorts. When checking the UPDRS-III files for these missing participants, we found the potential visit used by the authors, but these did not meet the criteria set to select the valid UPDRS-III scores (i.e. "PAG_NAME" was equal to "NUPDRS3" for these visits, but these were discarded when selecting only ON medication scores). For one participant missing in the Year 2 time point, we have not found any visit 2 years +/- 2 months after the Baseline visit. The visit selected for this participant was different in our cohort compared to the original authors cohort due to missing images, which could explain the reason for not finding back this participant for the Year 2 time point. Table 1 summarizes the cohort selection process.

Criterions	**N**
PPMI global query - Baseline	**102**
Participants belonging to the list provided by the authors at Baseline	**82**
Participants not belonging to the corresponding session list	4
Original session after the one obtained with PPMI query	2
Image of original session not available anymore in PPMI	2
PPMI global query - Year 1	**67**
Participants belonging to the list provided by the authors at Year 1	**51**
Participants not belonging to original list	2
PAG_NAME was NUPDRS3	2
PPMI global query - Year 2	**61**
Participants belonging to the list provided by the authors at Year 2	**41**
Participants not belonging to original list	4
PAG_NAME was NUPDRS3	3
Absence of corresponding score at follow-up time point	1
PPMI global query - Year 4	**46**
Participants belonging to the list provided by the authors at Year 4	**30**
Participants not belonging to original list	3
PAG_NAME was NUPDRS3	3

Table 1. Summary of cohort selection procedure. PPMI global query corresponds to the replication cohort, highlighted in **blue**. Participants belonging to the list provided by the authors composed the closest-to-original cohort, highlighted in **green**.

Image pre-processing

We downloaded functional images from the PPMI database manually for all participants selected in the replication cohort by using the image identifiers corresponding to the participants and visits selected. We also downloaded T1w images corresponding to the participants and visits selected in the replication cohort. If multiple T1w images were available for a participant at a given visit, we selected the one with the smallest identifier number (1st one in the meta-data table). Since imaging data from the PPMI online database were available in DICOM format, we converted them into the NIfTI format and we reorganized the dataset to follow the Brain Imaging Data Structure (BIDS) [28] (RRID:SCR_016124) using HeuDiConv v0.13.1 [29] (RRID:SCR_017427) on Docker v20.10.16.

Default reproduction pipeline

To pre-process the data, we started by building a pipeline reproducing the one described by the authors in [1]. The paper mentions that fMRI images were first realigned to the mean volume with affine transformations to correct for inter-volume head motion, using the MCFLIRT tool in the FSL toolbox [30] (RRID:SCR_002823). Then, images were brain-masked using AFNI 3dAutomask [31] (RRID:SCR_005927). Non-linear registration was performed directly to a common EPI template in MNI space using the Symmetric Normalization algorithm in ANTS [32] (RRID:SCR_004757). For denoising, motion-related regressors computed using ICA-AROMA [33] were concatenated with the nuisance regressors from affine head motion parameters computed with MCFLIRT and mean timeseries of white matter and cerebrospinal fluid. These nuisance signals were regressed out of the fMRI data in one step (i.e. all confounds concatenated in a single matrix and regressed from voxels timeseries).

Using this information, we reproduced the closest-possible pipeline to this description. We implemented this pipeline — referred to as the *default workflow* — using Nipype v1.8.6 (RRID:SCR_002502) [34], FSL v6.0.6.1, AFNI v23.3.01 and ANTs v2.3.4. We executed the pipeline with a custom-built Docker image available on Dockerhub https://hub.docker.com/repository/docker/elodiegermani/nguyen-etal-2021/general and built using NeuroDocker [35] with base image fedora:36 and a miniconda v23.5.2-0 [36] environment with Python v3.10. All pre-processing, feature computation and model training were run using custom Boutiques descriptors using Docker v20.10.16 and Boutiques v0.5.25 [37]. Boutiques descriptors for image processing and model training are available in Zenodo [38, 39].

In this *default reproduction workflow*, functional images were first realigned to the middle volume using FSL MCFLIRT, using affine registration (6 degrees of freedom), b-spline interpolation and mutual information cost function. The motion-corrected images were then skull-stripped using AFNI 3dAutomask with default parameters (clip level fraction of 0.5). Following this, ANTs symmetric normalization algorithm was used to normalize images to the MNI template. First, rigid, affine, and symmetric normalization transformations from native to MNI space were computed using the first volume of the brain-extracted functional images as source image and the MNI152NLin6Asym template, with a 2mm resolution as reference. The exact MNI template used for registration was not mentioned in the original paper. The choice of this particular template for our reproduction was due to the use of ICA-AROMA after registration. Indeed, to run ICA-AROMA in the MNI space or without FSL registration transform matrices, images must be in FSL's default MNI space, which is the MNI152NLin6Asym [40]. We downloaded this EPI template from C-PAC: https://github.com/FCP-INDI/C-PAC/blob/main/CPAC/resources/templates. We applied the computed transformations to functional images using ANTs also with B-Spline non linear registration.

For denoising, we regressed out several nuisance signals from the fMRI data, as in the original study. The 6 affine motion parameters computed using MCFLIRT were used as regressors. In addition, we ran ICA-AROMA v0.4.3-beta on data already registered in MNI space to extract motion-related components. All the components classified as motion-related were added as regressors to each participants.

For white-matter (WM) and cerebrospinal fluids (CSF) signals, the original paper did not contain any information about the method used by the authors to compute these signals. Thus, we implemented three different methods to try to reproduce the original workflow but also to compare the impact of pre-processing pipelines on the performance of the machine learning models and thus, the results of the study. In the *default workflow*, we chose to use AFNI to compute these regressors. We used the structural T1w images downloaded from PPMI and ran several analysis steps: brain extraction using 3dSkullstrip, segmentation using 3dSeg with defaults parameters, 3dCalc to extract the mask for WM and CSF, 3dResample to resample the masks to the functional image using nearest-neighbors interpolation and 3dMaskave to extract timeseries of voxels inside the WM and CSF masks. Then, we computed the mean timeseries across these voxels for WM and CSF and added these signals as nuisance regressors.

Variations of the reproduction pipeline

We also compared this workflow with two other methods to extract WM and CSF signals. The first method (pipeline *B.1 - default workflow with FSL segmentation*) used tools from FSL instead of AFNI to extract structural-derived masks. In this pipeline, BET was used to remove non-brain tissues from structural images, then the images were segmented using FAST to extract WM and CSF masks. The masks were resampled to functional images using affine registration implemented in FLIRT, and mean timeseries inside each mask were extracted using FSL's ImageMeants function in Nipype.

The second method (pipeline *B.2 - default workflow without structural priors*) did not involve image segmentation. We used mask templates available in FSL and Nilearn: MNI152_T1_2mm_VentricleMask from FSL for CSF, and WM brain-mask in MNI152 template resolution 2mm in Nilearn v0.10.2 [41] (RRID:SCR_001362) for WM. The masks were resampled to the functional images using a nearest neighbors interpolation in Nilearn, and mean timeseries inside each mask were also computed using Nilearn.

In all reproduction pipelines, the nuisance signals were regressed from the functional images in MNI space using FSL RegFilt. The denoised images were then used to compute the imaging features passed as input to the machine learning models.

Other pipelines variations

To explore the robustness of the original results to variations in the workflow, we also analyzed the functional and structural images using fMRIprep v23.0.2 [42] (RRID:SCR_016216), a robust pre-processing pipeline that requires minimal user input. We used default parameters for fMRIprep, except for the reference template that we set to MNI152NLin6Asym with a resolution of 2mm to be able to run ICA-AROMA afterwards [40].

Final preprocessed functional images in MNI space were then passed as input to ICA-AROMA to obtain motion-related components. The 6 motion regressors, WM and CSF mean timeseries extracted by fMRIprep were concatenated to the timeseries of the motion-related components identified by ICA-AROMA and regressed out from the pre-processed images using FSL RegFilt, as in the reproduction pipeline. This pipeline is referred to as *B.3 - fmriprep pipeline*.

Quality control

We implemented quality control checks at different steps of the pipelines. The purpose of these controls was to explore quality of data, but we did not exclude any participant due to data low quality, as this step was not performed in the original paper.

For each participant, we controlled the quality of functional pre-processing (motion correction, brain masking, and registration to MNI space) by superposing the pre-processed functional volume at each time point to an MNI-space brain mask, and visually inspecting a pre-defined image slice for incorrect registration or masking. We also visually inspected the 6 motion parameters identified during motion correction (rotation and translation in the x, y and z directions). We also computed the frame-wise displacement (FD) of head position as done in [43], calculated as the sum of the absolute volume-to-volume values of the 6 translational and rotational motion parameters converted to displacements on a 50 mm sphere (multiplied by $2 \times \pi \times 50$). We explored these values using the threshold used in [44] for the lenient strategy: identification of participants with mean FD > 0.55mm. Segmentations masks for WM and CSF obtained with the 2 different workflow variations were also visually inspected for failed segmentations. For the fMRIprep pipeline, we validated the quality of the processing using the log files produced by the pipeline, since these produce the same outputs as the quality control steps mentioned above.

Imaging features computation

In the original study, mean regional values of z-scored fALFF and ReHo were used as input features to the machine learning models, in addition to several clinical and demographic features. fALFF and ReHo were computed on the denoised fMRI data using C-PAC [45] (RRID:SCR_000862). Voxel-wise ReHo was computed using Kendall's coefficient of concordance between each voxel and its 27-voxel neighborhood. For ALFF and fALFF, linear de-trending and band-pass filtering were first applied to each voxel at 0.01–0.1 Hz, then the standard deviation of the signal was computed to obtain ALFF whole-brain maps. These maps were divided by the standard deviation of the unfiltered signal to obtain whole-brain fALFF maps. Z-scores maps for ReHo and fALFF were calculated at the participant-level.

For our reproduction, we used the original code used by the authors, available at https://github.com/DeepLearningForPrecisionHealthLab/Parkinson-Severity-rsfMRI/blob/master/ppmiutils/rsfmri.py. We followed the exact same steps as in the original paper to compute the raw ReHo and fALFF maps. However, a mask file was needed in the authors' code to compute the features. We thus applied AFNI 3dAutomask on the denoised fMRI data to obtain a brain mask for each participant.

The code shared by the authors did not include any z-scoring of the whole-brain maps for fALFF and ReHo, thus we used FSL's ImageMaths function to compute the z-score maps. Non z-scored maps (*C.1 - default workflow with no Z-scoring*) were also saved and set as input to the models for comparison. We also considered ALFF instead of fALFF as input measure (*C.2 - default workflow with ALFF*) as the authors also mentioned having tested this feature.

In the original paper, regional features were extracted from the ReHo and fALFF whole-brain maps using three different parcellations. These included the 100-ROI Schaefer [46] functional brain parcellation, modified with an additional 35 striatal and cerebellar ROIs, and the 197-ROI and 444-ROI versions of the Bootstrap Analysis of Stable Clusters (BASC) atlas [47]. These parcellations were used to compute the mean regional ReHo or fALFF values for each participant and performance of the machine learning models were compared between the parcellations. For the reproduction, we did not have access to the modified version of the Schaefer atlas used by the original

authors. Thus, we derived a similar custom atlas by using the 100-ROI Schaefer atlas available in Nilearn, the probabilistic cerebellar atlas available in FSL, from [48], and the Oxford-GSK-Imanova connectivity striatal atlas from [49], also available in FSL. The cerebellar and striatal atlases were respectively composed of 28 and 7 ROIs, which was consistent with the 35 ROIs mentioned in the original paper. We merged the ROIs from the Schaefer, cerebellar and striatal atlas in this order to build a custom 135-ROI atlas which we used to extract regional features.

The three atlases were resampled to the whole-brain ReHo and fALFF maps using Nilearn and a nearest-neighbor interpolation, as done by the authors. Mean regional values for each imaging feature and parcellation were also extracted using Nilearn.

Input features

Clinical and demographic features

In addition to imaging features, the authors used several clinical and demographic features as input to the machine-learning models. Clinical features included disease duration, symptom duration, dominant symptom side, Geriatric Depression Scale (GDS), Montreal Cognitive Assessment (MoCA), and presence of tremor, rigidity, or postural instability at Baseline. Baseline MDS-UPDRS score was also included as a feature when training models to predict outcomes at Year 1, Year 2, and Year 4. Demographic features included age, sex, ethnicity, race, handedness, and years of education.

We searched for the mentioned input features using the study files in the PPMI database, as done by the authors (see https://github.com/DeepLearningForPrecisionHealthLab/Parkinson-Severity-rsfMRI/blob/master/ppmiutils/dataset.py). For each feature, we searched for the corresponding columns in the study files and used the same character encoding method as the authors. The different features used and the methods to search and encode them for input to the models are shown in Supplementary Table S1.

To evaluate the robustness of the findings to different analytical conditions, we also compared the results obtained with different sets of features. In pipeline *D.4 - default workflow with only imaging features*, we trained models using only imaging features (regional measures of fALFF and ReHo), i.e., without clinical or demographic features. In pipeline *D.3 - default workflow with no imaging features*, we removed imaging features and trained models only on clinical and demographic features. Following an update of the PPMI database, the feature for dominant disease side was deprecated and only available as an archive file in the version of the database we had access to. We included the feature in the *default workflow* and removed it in another variation workflow, to assess the impact of this feature (*D.1 - default workflow with no dominant disease side*). For models trained to predict MDS-UPDRS scores at Year 1, Year 2, and Year 4, Baseline MDS-UPDRS score was included as feature. However, due to the potential large effect of including this variable on the results, we trained a model with all features except this one and compared the performance of prediction models with and without the feature (*D.2 - default workflow with no Baseline MDS-UPDRS*).

Outcome measurement

In [1], the authors used the above-mentioned imaging, clinical, and demographic features to predict MDS-UPDRS total scores. The MDS-UPDRS score consists of 4 parts with 51 items, each item values from 0 to 5. To compute the total scores, we summed the values of the 4 different parts available in PPMI study files. We used: MDS-UPDRS part Ia entered by a rater (PPMI column "NP1RTOT"), part Ib for the patient questionnaire (column "NP1PTOT"), part II ("NP2TOT"), part III ("NP3TOT") and part IV ("NP4TOT").

Missing values in "NP4TOT" columns were replaced with zeros, as done by the authors. There were no participants with missing values for the other parts of the score.

Model selection and performance evaluation

We trained and optimized separate machine learning models to predict MDS-UPDRS scores from either ReHo or fALFF features, along with clinical and demographic features. Four machine learning models architectures were implemented using scikit-learn v1.3.0 [41] and were tested for each target-imaging feature (fALFF or ReHo) combination: ElasticNet regression, Support Vector Machine (SVM) with a linear kernel, Random Forest with a decision tree kernel, and Gradient Boosting with a decision tree kernel. Each parcellation was also implemented, which resulted in 12 different combinations of model and parcellation per imaging feature and time point.

For hyperparameter optimization (1) and performance estimation (2), authors used a nested cross-validation scheme, i.e., each model architecture × hyperparameter × parcellation combination was evaluated using (1) a 10-fold cross-validation inner-loop applied to the n-1 participants in the cohort and from which the combination with the lowest root mean squared error (RMSE) was selected, (2) a leave-one-out (LOO) cross-validation outer-loop where each iteration trained the selected model on all the participants in the cohort except one, and tested the model on the remaining held-out participant. To evaluate the impact of the evaluation pipeline on the results, we implemented a different nested cross-validation loop for model selection and evaluation for the *default workflow*. Figure 2 illustrates the different methods implemented. We evaluated the performance of each combination of model × parcellation separately: the 10-fold cross-validation inner-loop was used to select the set of hyperparameters (e.g. maximum tree depth for Random Forests) with the lowest RMSE, this set was used to train a model on all except one participants in the outer-loop and we tested the model on the held-out participant. Thus, we obtained performance estimates for each model × parcellation combination.

We also reported results obtained using the exact nested cross-validation scheme explained in the paper (*E.1 - Workflow with paper's nested cross-validation*), i.e., the performance on each outer-fold is assessed with the best model × hyperparameter × parcellation combination found on the 10-fold cross-validation of the inner-loop and averaged across outer-folds. Finally, as authors reported only the best performing model and parcellation for each imaging feature type and time point, we also reported the results we would have obtained had we only used the best model and parcellation reported in the paper (*E.2 - Workflow with only paper's best model reporting*).

All models were trained using scikit-learn, we used the set of hyperparameters available in the authors code to train and optimize the models.

Evaluation metrics

As in the original paper, performance metrics included the coefficient of determination (R2), which represents the percentage of variance explained by the model, and the root mean squared error (RMSE), as implemented in scikit-learn.

We defined a null performance to compare our R2 values to using permutation test. We fixed the model and parcellation scheme with ElasticNet and Schaefer atlas. We ran 1000 permutations on the target labels and obtained performance for each feature and timepoint. At each permutation, we performed a nested cross-validation with 5-folds cross-validation as inner-loop and outer-loop. We optimized the hyper-parameter set of the model as done with the "real" models in the inner-loop and evaluated performance on the outer-loop. R2 values obtained using the different workflows were compared to

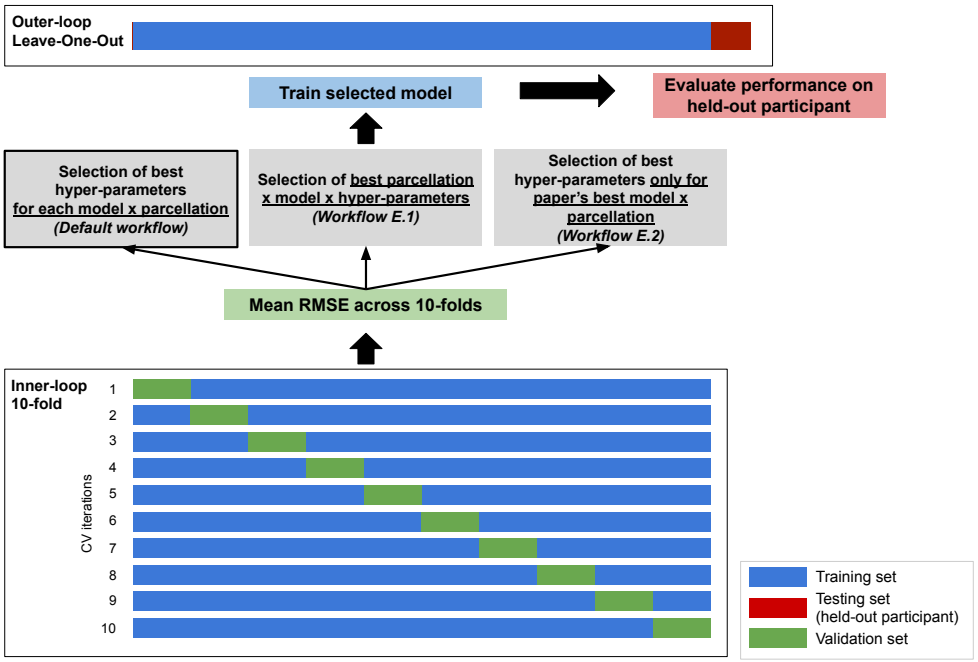

Figure 2. Workflow of model selection and performance evaluation. This workflow represents one iteration of the outer-loop with Leave-One-Out cross-validation and is iterated over all the dataset to estimate mean performance.

this null performance to check if the models did not learn to predict only the average value.

To evaluate the models' ability to classify high versus low severity participants, a threshold was set to separate the participants and each model's predictions were thresholded post-hoc. This threshold was computed by using the average of the median MDS-UPDRS score at each of the four time points. In [1], the threshold was 35. We computed this threshold the same way for the replication cohort and for the closest-to-original cohort. We obtained a value of 36 for the replication cohort and 35 for the closest-to-original one. Authors also mentioned having found no significant difference ($p > 0.05$) between the high and low-severity groups in motor predominance (Part III score as a percentage of total score) at each time point. With our thresholds, we ran two sample t-tests between high and low severity groups in the two cohort and did not find any significant difference with $\alpha = 0.05$ either in any cohort or time point. Performance metrics for this secondary classification outcome included area under the receiver operating characteristic curve (AUC), positive predictive value (PPV), negative predictive value (NPV), specificity, and sensitivity.

Feature importance

As in [1], we measured feature importance in the models trained for each time point and imaging feature (fALFF or ReHo). For the ElasticNet and SVM models, we used the coefficients of the trained models to determine feature importance, since coefficients of higher magnitude indicate more important features in these two models. The sign of the coefficient was indicative of whether the feature was positively or negatively associated with the prediction target. For Random Forest and Gradient Boosting models, we used impurity-based feature importance coupled with univariate linear correlation to determine the direction of the association. Feature importance was computed on each

iteration of the outer-loop and the median importance was reported for each feature.

To name the imaging features, we used the same method as the authors of [1]: the centroid of each feature's ROI was computed, if the feature was located in a ROI of the Automated Anatomical Labeling (AAL) atlas [50], this label was allocated to the ROI. If not, we searched for the nearest ROI of the AAL atlas.

Results

Cohort selection

Using the method described above, we built two cohorts from the PPMI database: the replication cohort and the closest-to-original cohort.

Table 2 shows the demographics and Baseline clinical characteristics of the replication and closest-to-original cohorts compared to the original cohort reported in [1]. The replication cohort was composed of respectively 102, 67, 61 and 46 participants for time points Baseline, Year 1, Year 2, and Year 4. The closest-to-original cohorts at the same time points were composed of respectively 82, 51, 41 and 30 participants.

Compared to the original cohort, our replication cohort showed similar demographics characteristics at each time point, except at Year 4 where our replication cohort showed a significantly higher age on average than in the original cohort ($p < 0.01$). Regarding clinical variables, mean MoCA score, GDS total score and Hoehn-Yahr stage were similar between the two cohorts at all time points. However, we found higher mean disease durations in the replication cohort than in the original one at all time points, for instance at Baseline with (866.9 days $\pm$ 598.7 days) in replication vs (770 days $\pm$ 565 days) in original. This difference was not significant at threshold $p < 0.05$. We also observed lower mean MDS-UPDRS scores at Baseline in the replication cohort for all time points except Baseline, with significant difference at Year 2 ($p < 0.05$) only. For these two time points, even if mean Baseline scores in the replication cohort significantly differed from the original ones, mean MDS-UPDRS scores at prediction time point were more similar to the original one. At Year 4, however, we also found a higher mean MDS-UPDRS score at prediction time point than in the original cohort, but this difference was not significant at $p < 0.05$.

The closest-to-original cohort exhibited almost the same characteristics as the original one at Baseline. For subsequent time points, we found some differences, in particular at Year 2 and at Year 4: participants were older in the closest-to-original cohort than in the original study at Year 4 ($p < 0.05$), Baseline mean MDS-UPDRS score was lower (significant for Year 2 and Year 4 at $p < 0.05$ and $p < 0.01$ respectively) and mean MDS-UPDRS score at prediction time point was similar to the original cohort except at Year 4.

These differences for the Year 1, Year 2, and Year 4 cohorts could be related to the evolution of the PPMI database in which sessions were added and removed since the authors queried it for the original study. For these time points, we were not able to find back all the participants that were included in the original cohort: the patients included in our closest-to-original cohorts represented respectively 96% (Year 1), 91% (Year 2) and 91% (Year 4) of the patients included in the original cohort. However, only represented 76% (Year 1), 67% (Year 2), and 65% (Year 4) of the replication cohort was composed of patients of the original cohort.

Figure 3 compares the distribution of MDS-UPDRS scores in our cohorts with the one in the original cohort reported in Figure S1 in [1]. Distributions of MDS-UPDRS scores at Baseline were similar between our two cohorts but seemed different from the original cohort one. The observed difference between the original and closest-to-original distributions might result from differences in MDS-UPDRS score calculations, or from

	Baseline			Year 1			Year 2			Year 4		
	Orig.	Repro.	Closest	Orig.	Repro.	Closest	Orig.	Repro.	Closest	Orig.	Repro.	Closest
% Caucasian	95.1	95.1	93.9	94.4	94.0	94.1	97.8	95.1	95.1	97.0	97.8	96.7
% African-American	2.4	2.0	2.4	1.9	1.5	0.0	0	1.6	0.0	0	0.0	0.0
% Asian	3.7	2.9	3.7	5.6	4.5	5.9	4.4	3.3	4.9	3.0	2.2	3.3
% Hispanic	1.2	1.0	0.0	0	1.5	0.0	0	1.6	0.0	0	0.0	0.0
% Male	67.0	66.7	67.1	68.5	65.7	68.6	82.2	80.3	85.4	75.8	67.4	73.3
% right-handed	89.0	89.2	89.0	85.2	85.1	84.3	88.9	90.2	90.2	87.9	84.8	86.7
Mean age, years	62.1 ± 9.8	62.0 ± 9.5	62.1 ± 9.7	61.9 ± 10.3	62.2 ± 9.9	63.0 ± 10.4	63.6 ± 9.2	64.7 ± 9.1	65.9 ± 9.4	59.5 ± 11.0	**66.2 ± 10.1****	**63.8 ± 11.0***
Mean years of education	15.6 ± 3.0	15.6 ± 2.8	15.7 ± 2.9	15.1 ± 3.2	15.5 ± 2.9	15.4 ± 2.9	15.1 ± 3.3	15.4 ± 2.8	15.5 ± 3.0	15.0 ± 3.4	15.3 ± 3.0	15.2 ± 3.4
Mean disease duration at Baseline, days	770 ± 565	866.9 ± 598.7	760.3 ± 559.2	808 ± 576	904.1 ± 614.5	808.5 ± 580.0	771 ± 506	867.5 ± 516.3	732.0 ± 462.8	532 ± 346	746.6 ± 624.6	464.6 ± 294.9
Mean MDS-UPDRS at Baseline	33.9 ± 15.8	34.5 ± 15.6	33.9 ± 16.1	38.0 ± 20.9	33.4 ± 15.1	34.1 ± 15.4	40.2 ± 18.2	**35.0 ± 15.1***	**35.2 ± 16.1***	34.9 ± 15.7	30.7 ± 13.9	**26.1 ± 11.4****
Mean MDS-UPDRS at timepoint	-	-	-	39.2 ± 21.6	40.7 ± 24.5	39.9 ± 22.0	40.9 ± 18.5	40.0 ± 18.7	40.7 ± 18.7	35.9 ± 16.5	41.5 ± 19.8	34.2 ± 16.2
Mean MoCA at Baseline	26.7 ± 2.8	26.5 ± 3.0	26.4 ± 2.8	26.9 ± 3.2	27.0 ± 2.9	26.7 ± 3.1	26.7 ± 3.5	27.0 ± 2.5	26.5 ± 2.4	27.5 ± 2.3	26.8 ± 3.2	27.4 ± 2.6
Mean GDS at Baseline	5.4 ± 1.4	5.4 ± 1.4	5.4 ± 1.5	5.4 ± 1.6	5.5 ± 1.8	5.5 ± 1.9	5.4 ± 1.2	5.5 ± 1.3	5.6 ± 1.3	5.4 ± 1.7	5.8 ± 1.8	5.6 ± 1.7
Mean Hoehn-Yahr stage	1.8 ± 0.5	1.7 ± 0.5	1.7 ± 0.5	1.8 ± 0.5	1.8 ± 0.6	1.7 ± 0.5	1.8 ± 0.5	1.9 ± 0.5	1.9 ± 0.5	1.7 ± 0.5	1.9 ± 0.5*	1.8 ± 0.5
Number of subject	82	102	82	53	67	51	45	61	41	33	46	30

Table 2. Demographic and clinical variables for the different cohorts. Orig. = original paper cohort. Repli. = replication cohort. Closest = closest-to-original cohort. Values are reported in percentages of the cohort or in mean values ± standard deviation. Significance testing was performed using two sample t-test between the original cohort and the replication and closest cohort respectively. Bold text represent features showing a significant difference, * represent significance at $p < 0.05$ and ** at $p < 0.01$.

the fact that different sessions were used for 4 of the participants in the closest-to-original cohort compared to the original one. At Year 1, however, the closest-to-original cohort presented a MDS-UPDRS score distribution more similar to the original one than the replication one, suggesting that the differences at Baseline did not originate in differences in MDS-UPDRS score calculations. We found no significant difference between the distribution of MDS-UPDRS scores in the replication and closest-to-original cohort neither at Baseline nor at Year 1 using Kolmogorov-Smirnov distribution testing.

Image quality control

After running the pre-processing pipelines, we checked the resulting images and looked for potential pipeline failures. Regarding registration, all participants brains were correctly registered to the MNI space after visual inspection. Brain masking was also successful for most of the participants, except for 2 in which we found a small artifact in the inter-hemispheric area. Given the low magnitude of this artefact and its location, we decided to keep these two participants in the study.

Most participants of the study showed high movement parameters. Indeed, out of 102, 80 showed at least one time point with a frame-wise displacement superior to 0.5mm. However, since the authors in [1] did not remove high-motion volumes within participants, and that completely removing participants with high-motion volumes would highly decrease our cohort's sample size, we chose to keep all participants and all volumes.

Regarding segmentation masks, after visual inspection no significant artifact was found for any participants using AFNI segmentation in default workflow. For some participants, small distortions were found in particular close to brain extremities (inter-hemispheric area or close to the skull in occipital and parietal regions). Using FSL segmentation however, we found brain masking issues that had impacts on segmentation quality. We used BET using default parameters to skullstrip images before segmentation and since we chose to explore the impact of different default implementations of pipelines, we did not exclude the segmentations for any participant nor segmentation workflow.

With the fMRIprep pipeline, observations were similar regarding movement parameters and registration. There was no large artefact in the segmentation masks.

Performance of the *default workflow*

The first objective of this study was to reproduce the models described in [1] and to compare their performance with the one in the original study. In our default workflow, we implemented the default choices described in Figure 1: closest-to-original cohort, image pre-processing pipeline with AFNI segmentation, z-scoring of whole-brain fALFF and ReHo maps, use of all demographic, clinical and imaging features described in the original paper, and the model selection method derived from the authors' code.

We trained 12 models per time point (Baseline, Year 1, Year 2, Year 4) and imaging feature (fALFF or ReHo), corresponding to 4 machine learning models $\times$ 3 brain parcellations. We reported for each imaging feature and time point the performance of the 12 models in Table 3.

Chance levels were computed using permutation tests as described in the Evaluation metrics section. We obtained R2 values that represented the chance prediction performance at different time point for fALFF and ReHo. These values are also presented in Table 3.

Using the default workflow, we obtained different performance for all models $\times$ parcellation combination. The best performance across these combination was different from those reported in [1] at all time point. At Baseline, our best model performed better than chance but we did not manage to reach a R2 value close to the one reported in the original paper with any model or feature. Moreover, the best-performing models were

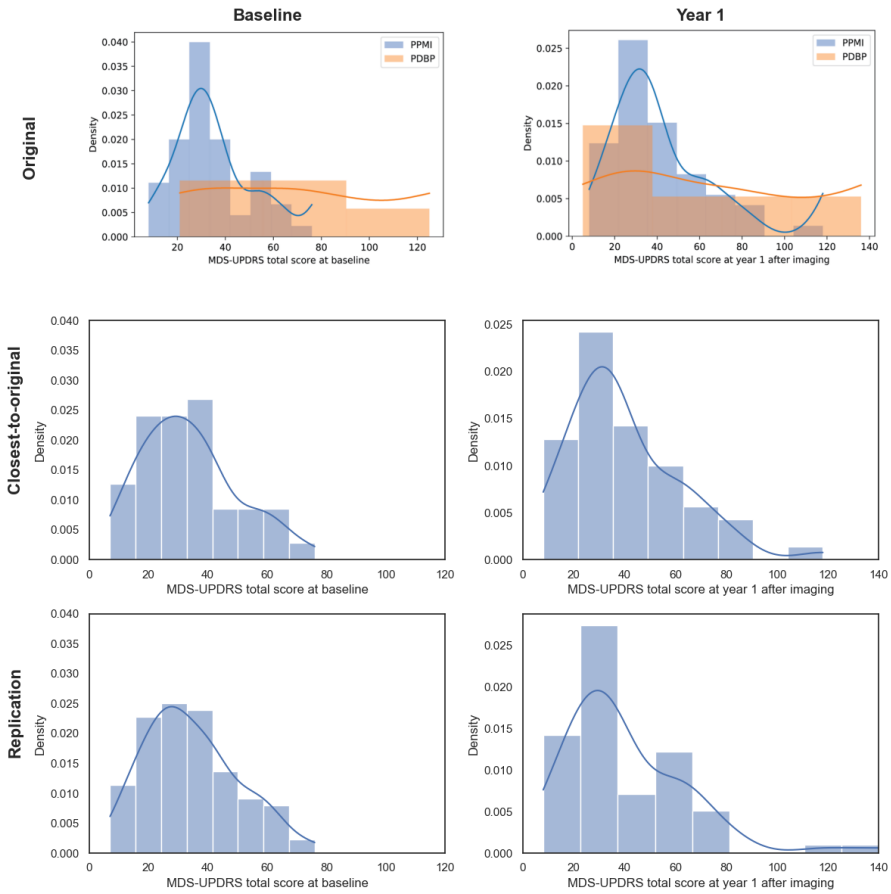

Figure 3. Distribution of MDS-UPDRS scores reported in the original paper's cohort (*top*: Figure S1 extracted from [1]), the replication cohort (*middle*) and the closest-to-original cohort (*bottom*).

different from those reported in the original study: instead of Schaefer atlas and Gradient Boosting for both fALFF and ReHo features, we found for fALFF the Gradient Boosting Regressor with BASC197 atlas, with R2=0.205 (original R2=0.242) and ElasticNet and Schaefer for ReHo with R2=0.124 (original R2=0.304).

At Year 1, the performance of our models was better than reported in the original study, with an increase of the R2 of 28% and 18% for fALFF and ReHo respectively. For other time points (Year 2 and Year 4), results were also different from those reported in [1]. These differences were not constant between ReHo and fALFF at Year 2, but were similar at Year 4: for fALFF, we obtained higher R2 scores than in the original study at Year 2 and at Year 4 (0.529 and 0.397 compared to 0.463 and 0.152 in the original paper); for ReHo, we obtained lower R2 scores than in the original ones at Year 2 (0.344 instead of 0.471) and higher R2 scores at Year 4 (0.312 compared to 0.255 in the original study). For these two time points, the mean MDS-UPDRS scores at Baseline were significantly different between the original cohort and our closest-to-original cohort, which might explain these differences in performance.

At each time point, the best model x parcellation combination performed better than chance-level. Some of the combinations led to very low performance, for instance SVM with Schaefer atlas at Year 2. At every time point and with every feature (except at Year 1 with fALFF), at least one combination gave a performance lower than chance. Average performance across model and parcellation combinations were sometimes very

close to chance-level. This highlight the importance of model selection and performance reporting.

Robustness to workflow variations

We assessed the performance of the different models for each time point and feature for different variations of the analysis workflow (Figure 4).

Workflow A.2, in which we trained the different models on the replication cohort instead of the closest-to-original one, showed only small differences in R2 values with the *default workflow*, except for fALFF at Year 1 and ReHo at Year 4. Indeed, performance was slightly lower at Year 1 for fALFF and higher at Year 4 for ReHo, with raw effect size above 0.15. At Year 1, the replication cohort was composed of 16 more participants than the closest-to-original cohort and exhibited a lower mean MDS-UPDRS score at Baseline compared to the original cohort. At Year 4, we also found differences in term of sample size, age of participants and Baseline MDS-UPDRS score between the replication cohort, the original one and the closest-to-original one. These differences might explain the variations between models performance, even if R2 values remained better-than-chance for Year 1 and close to other performance obtained with different variations. Best model performance of *workflow A.2* remained better than chance-level.

Performance of models trained with variations in pre-processing pipeline (*workflows B.1, B.2 and B.3*) was similar to those of the default workflow, with R2 absolute difference with the *default workflow* below 0.15 except at Year 4 with fALFF in which the *B.2 workflow* (no structural segmentation) led to lower R2 values and at baseline with fMRIprep pipeline (*B.3 workflow*). For these, best performance was better than chance but not the average performance across model and parcellation combination.

Regarding the impact of feature computation variations (*workflows C.1 and C.2*), we found better performance at Baseline for workflows *C.2 - default workflow with ALFF* in which the best model $\times$ parcellation combination led to a better R2 value than the one reported in the original study (0.325 vs 0.242 in the original paper). We also observed this phenomenon with the *C.1 workflow* in which we used non z-scored ReHo maps: we found a higher performance than the one obtained with the default workflow and reported in the original study ($R2 = 0.374$). For these two variations, R2 differences with default remained lower than 0.1. At Year 1 and Year 4 with fALFF however, the use of ALFF instead of fALFF (*workflow C.2*) led to lower performance (R2 mean absolute difference above 0.15 and average performance below chance level). This observation was not found at Year 2.

For Year 1 and Year 2 predictions, the set of input features (*workflows D.*) had a large impact on the performance of these models. In particular, models trained without Baseline MDS-UPDRS score (D.2) and with only imaging features (D.4) showed lower R2 values with for fALFF and for ReHo at Year 1 and Year 2 (R2 absolute difference above 0.2), which suggests that Baseline MDS-UPDRS played a central role in the prediction of MDS-UPDRS at follow-up visits compared to imaging features. It also explains why variations in the extraction of imaging features (pre-processing or computation) only had a lower impact on the performance for these two time points.

Overall, at Year 1 and Year 2, performance seemed to be driven mostly by clinical and demographic features, in particular by MDS-UPDRS Baseline scores. At Baseline and Year 4, other variations related to image features (pre-processing and feature computation) were associated with larger changes in performance. For all workflows, time points and feature, best performing model x parcellation combination always exhibited better than chance performance.

Time	Feature	Type	ElasticNet			SVM			GradientBoosting			RandomForest		
			schaefer	basc197	basc444	schaefer	basc197	basc444	schaefer	basc197	basc444	schaefer	basc197	basc444
Baseline	fALFF	Orig.							**0.242**					
		Repli.	0.04	-0.035	-0.045	-0.718	-0.241	-0.182	-0.039	**0.205**	0.061	-0.024	0.068	0.02
		Null	**-0.041**											
	ReHo	Orig.							**0.304**					
		Repli.	**0.124**	0.057	0.117	-0.3	-0.4	-0.152	-0.102	0.028	0.027	0.024	0.022	0.099
		Null	**-0.036**											
Year 1	fALFF	Orig.	**0.558**											
		Repli.	0.453	**0.717**	0.5	0.519	0.216	0.185	0.622	0.575	0.506	0.369	0.499	0.444
		Null	**-0.079**											
	ReHo	Orig.	**0.453**											
		Repli.	**0.535**	0.434	0.512	0.04	-0.094	-0.01	0.36	0.261	0.289	0.442	0.392	0.393
		Null	**-0.077**											
Year 2	fALFF	Orig.	**0.463**											
		Repli.	**0.529**	0.277	0.285	-0.031	0.108	-0.413	-0.19	0.08	0.01	0.138	0.206	0.09
		Null	**-0.101**											
	ReHo	Orig.	**0.471**											
		Repli.	**0.344**	0.191	0.287	-0.915	-0.741	-0.051	-0.03	0.001	-0.033	0.267	0.121	0.251
		Null	**-0.094**											
Year 4	fALFF	Orig.				**0.152**								
		Repli.	**0.397**	0.115	0.351	0.196	-0.134	-0.296	0.08	0.411	-0.355	0.079	0.338	0.01
		Null	**-0.129**											
	ReHo	Orig.				**0.255**								
		Repli.	0.072	0.09	-0.175	-0.12	-0.23	-0.139	-0.017	**0.312**	0.041	-0.007	0.02	0.0
		Null	**-0.141**											

Table 3. Predictive performance achieved for each MDS-UPDRS time point and each imaging feature type, computed through leave-one-out cross-validation. Metric: R2, coefficient of determination. Green text corresponds to original performance reported in [1]; Blue text corresponds to best performance achieved during replication; Red text corresponds to chance level computed using permutation test.

Model choice and performance reporting

Table 4 compares the results obtained using different model selection and evaluation methods. Using the nested cross-validation described in the paper (*Workflow E.1*), we obtained lower results than the original ones and than the ones obtained with our best models for all time points (for instance, $R2 = 0.049 vs 0.205$ with our best model for prediction with fALFF at Baseline). Using this method, the models at Year 1 and Year 2 were still well performing compared to other time point, for both ReHo and fALFF, with particularly high R2 values (between around 0.4 and 0.6) obtained using any reporting method.

Results computed using the same model and parcellation as the best performing combinations in the original paper (Table 2 from [1]) (*Workflow E.2*) also had lower performance than in the original study, for all time points (e.g. $R = -0.102$ for prediction with ReHo at Baseline). However, as observed for nested cross-validation, the performance obtained with these models at Year 1 and Year 2 was still high and close to the ones obtained with our best models. We speculate that the effect size detected with models at these time points was large and thus, tended to be more reproducible across optimization schemes.

In [1], authors also report the model's ability to classify high- versus low-future severity subjects. The performance obtained for this task was consistent with the observation made on R2 values: models with high performance in terms of R2 were usually good at distinguishing high and low severity patients (e.g., AUC of 0.805 and 0.767 for prediction at Year 1 with respectively fALFF and ReHo using the *default workflow*).

Feature importance

To further explore the reproducibility and replicability of findings in [1], we measured feature importance for the ReHo and fALFF imaging features and the default reproduction workflow, across all time points. Figure 5 and 6 compare the feature importances obtained with the *default workflow* to the ones reported in the original study.

Feature importance showed relatively few overlap between the ones obtained using our models and those reported in the original study, especially for imaging features, at all time points. For instance, for fALFF at Baseline, the left postcentral region was identified as the most important feature for prediction in our study and was not identified in the original study. For ReHo, we found no important imaging feature that was similar to the ones detected in the original study. However, for some brain regions for which an imaging feature was identified as an important feature, hemispheric opposites or sub-parts of the same global regions were identified in our models compared to the original detected features. For instance, the middle cingulum was identified in our Baseline model with ReHo but in the left hemisphere instead of the right one in the original paper. For this model, regions of the frontal cortex were also detected as important in the original paper, but those we found were very close or were part of the same lobe/region (e.g. frontal supero-orbital and middle in original, frontal inferior in ours). Regions identified for fALFF and ReHo were also different at Baseline, consistently with the findings of [1].

For other time points, the main feature of importance was the Baseline MDS-UPDRS score for both fALFF and ReHo and other features had a lower importance value, in particular at Year 1 and at Year 2. This observation was also supported by the performance of models that did not include the Baseline MDS-UPDRS score in their feature set: these models showed lower performance at these two time points compared to the default models ($p < 0.01$).

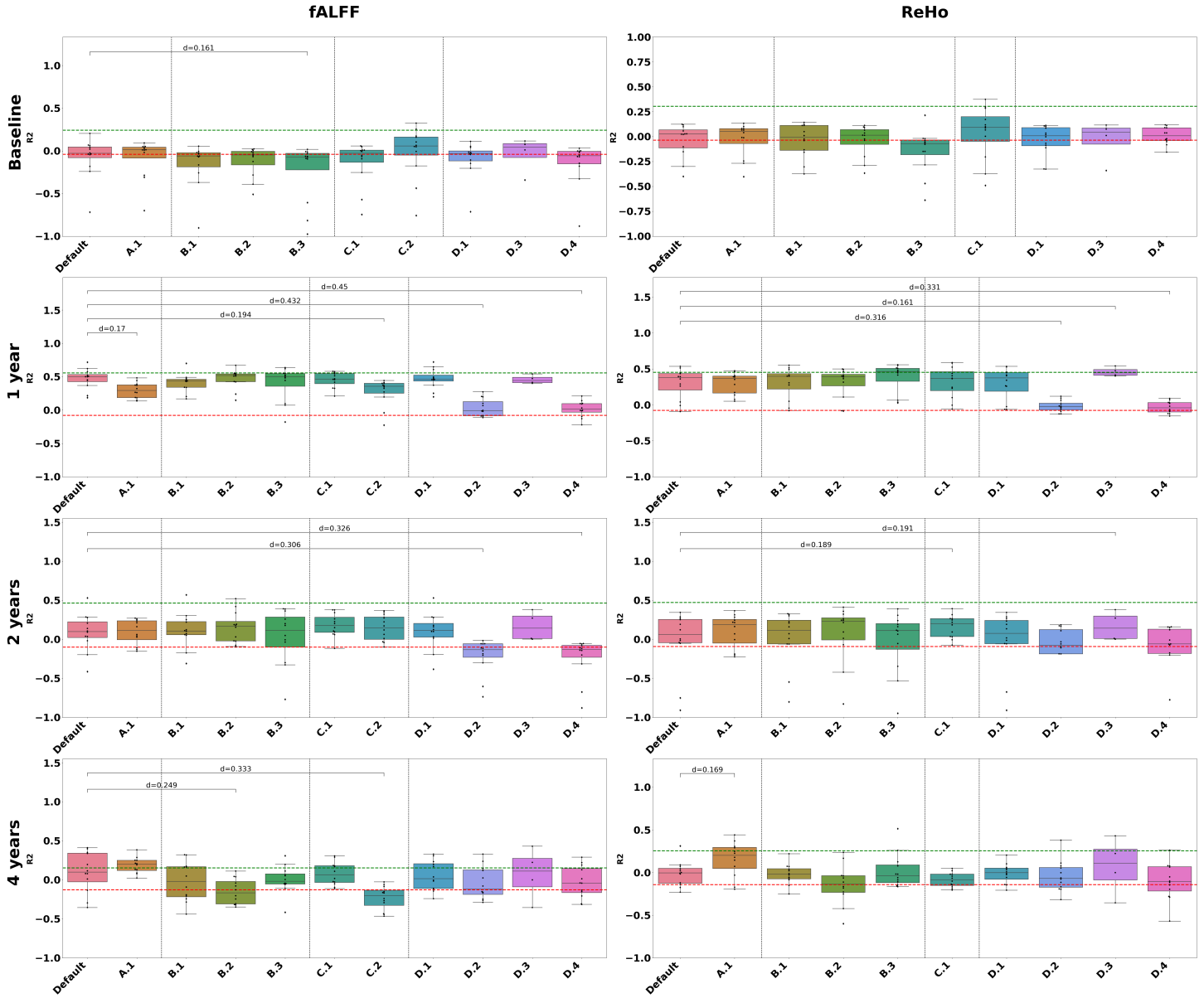

Figure 4. Performance of models trained for prediction at each time point, using fALFF or ReHo, with variations in the workflow. Boxes represent the performance (R2 values) of the 12 models (4 models × 3 parcellations). Black horizontal dashed lines show the R2 value reported in the original study for the corresponding time point and feature. Red horizontal dashed lines show the chance-level computed using permutation test. Raw effect sizes (d) are computed as absolute difference between the mean R2 performance with *default workflow* and mean R2 performance with other variations. Only large differences (above threshold $d = 0.15$) are reported.

- *Workflow A.1* - default workflow with replication cohort.
- *Workflow B.1* - default workflow with FSL segmentation,
- *Workflow B.2* - default workflow without structural priors,
- *Workflow B.3* - fMRIprep pipeline.
- *Workflow C.1* - default workflow with no Z-scoring,
- *Workflow C.2* - default workflow with ALFF.
- *Workflow D.1* - default workflow with no dominant disease side,
- *Workflow D.2* - default workflow with no Baseline MDS-UPDRS,
- *Workflow D.3* - default workflow with no imaging features,
- *Workflow D.4* - default workflow with only imaging features.
- *Workflow E.1* - default workflow with paper's nested cross-validation,
- *Workflow E.2* - default workflow with only paper's best model reporting.

Time point	Feature	Type	R2	RMSE	AUC	PPV	NPV	Spec.	Sens.
Baseline	fALFF	Original	0.242	14.006	0.668	60.0%	74.0%	75.5%	58.1%
		Default	0.205	14.26	0.584	51.7%	66.0%	71.4%	45.5%
		Workflow E.1	0.049	15.6	0.514	42.3%	60.7%	69.4%	33.3%
		Workflow E.2	-0.039	16.31	0.493	39.4%	59.2%	59.2%	39.4%
	ReHo	Original	0.304	13.415	0.674	59.4%	75.0%	73.5%	61.3%
		Default	0.124	14.98	0.716	63.9%	78.3%	73.5%	69.7%
		Workflow E.1	-0.164	17.26	0.528	43.8%	62.0%	63.3%	42.4%
		Workflow E.2	-0.102	16.8	0.493	39.3%	59.3%	65.3%	33.3%
Year 1	fALFF	Original	0.558	14.256	0.753	70.4%	80.0%	71.4%	79.2%
		Default	0.717	11.6	0.805	75.9%	86.4%	73.1%	88.0%
		Workflow E.1	0.569	14.3	0.786	73.3%	85.7%	69.2%	88.0%
		Workflow E.2	0.453	16.11	0.69	62.9%	81.2%	50.0%	88.0%
	ReHo	Original	0.453	15.861	0.753	70.4%	80.0%	71.4%	79.2%
		Default	0.535	14.85	0.767	71.0%	85.0%	65.4%	88.0%
		Workflow E.1	0.483	15.67	0.726	70.4%	75.0%	69.2%	76.0%
		Workflow E.2	0.535	14.85	0.767	71.0%	85.0%	65.4%	88.0%
Year 2	fALFF	Original	0.463	13.426	0.765	78.6%	76.5%	68.4%	84.6%
		Default	0.529	12.68	0.669	69.2%	66.7%	55.6%	78.3%
		Workflow E.1	0.478	13.35	0.669	69.2%	66.7%	55.6%	78.3%
		Workflow E.2	0.529	12.68	0.669	69.2%	66.7%	55.6%	78.3%
	ReHo	Original	0.471	13.322	0.739	75.9%	75.0%	63.2%	84.6%
		Default	0.344	14.95	0.635	65.5%	66.7%	44.4%	82.6%
		Workflow E.1	0.272	15.76	0.607	63.3%	63.6%	38.9%	82.6%
		Workflow E.2	0.344	14.95	0.635	65.5%	66.7%	44.4%	82.6%
Year 4	fALFF	Original	0.152	14.957	0.636	64.7%	62.5%	62.5%	64.7%
		Default	0.411	12.19	0.833	91.7%	77.8%	93.3%	73.3%
		Workflow E.1	0.242	13.83	0.733	73.3%	73.3%	73.3%	73.3%
		Workflow E.2	-0.134	16.92	0.633	66.7%	61.1%	73.3%	53.3%
	ReHo	Original	0.255	14.015	0.699	73.3%	66.7%	75.0%	64.7%
		Default	0.312	13.18	0.667	72.7%	63.2%	80.0%	53.3%
		Workflow E.1	-0.044	16.23	0.567	60.0%	55.0%	73.3%	40.0%
		Workflow E.2	-0.23	17.62	0.6	63.6%	57.9%	73.3%	46.7%

Table 4. Performance reported using different model selection and evaluation methods. "Original" is the performance reported in the Original study [1]. "Default" is the performance obtained with the model × parcellation that obtained the best performance during reproduction. "Workflow E.1" is the performance obtained when using the nested cross-validation scheme described in the paper (i.e. optimizing model × parcellation in the inner fold). "Workflow E.2" is the performance obtained with the model and parcellation reported in the paper.

Discussion

Summary

We investigated the reproducibility and replicability of the predictive models of PD progression described in [1]. Using the *default reproduction workflow*, i.e., with methods and cohorts closest to the ones described in [1], the performance of our best models was better than chance ($R2 > 0$). However, the models performance were in some cases slightly different from the ones reported in the original study. For both ReHo and fALFF, we found lower performance than the one reported in the original study at Baseline with our *default workflow*. This performance was also higher than in the original study at Year 1, Year 2 and Year 4. These values remained however close to those reported in the original study. Thus, using a cohort and methods adapted from [1], we were able to train several machine learning models that predicted Parkinson's disease progression (MDS-UPDRS scores at Baseline, Year 1, Year 2, and Year 4) with a performance higher than chance and with values comparable to those reported in the original study for most models. On these criteria, we could conclude that the replication experiment was successful.

Taking into account other criteria than model performance comparison, the success of the reproduction is less clear. For instance, we could not compare our reproduction workflow with the original one, neither with code (in particular for the pre-processing) nor derivatives data such as pre-processed images or whole-brain fALFF or ReHo maps. Thus, we were not able to verify if we correctly reproduced this pipeline, in either of our variations. Moreover, we were not able to retrieve the features identified as important by the authors during feature importance evaluation. This step was complex to reproduce since our best performing model x parcellation combination did not match the ones reported in the original paper at several time points, which questions the comparability of the features. In addition, we did not have access to the exact same atlas used in the original paper for the Schaefer parcellation, and we also did not have access to the labels used by the authors to name the regions. We attempted to reproduce the region names using the method described in [1], but we could not be certain that our regions matched the original ones.

When introducing specific variations in the workflow, we managed to obtain results that were more similar to the original ones, in particular when changing the feature computation method at Baseline. Some changes in the *default workflow* also led to lower performance, for instance at Year 1 and at Year 2 when removing Baseline MDS-UPDRS score or when using only imaging features. For these time points in particular, variations of the pre-processing pipeline (workflows B.), feature computation (workflows C.) and model choice and reporting (workflows E.) had little impact on the performance of the models compared to other time points. We speculate that imaging features only were of low importance in the models prediction for these time points compared to other time points (Baseline and Year 4) for which variations on image computation (pre-processing or feature) had a larger impact. Without variations (i.e. with the *default workflow*), performance of models at Baseline and Year 4 time points was already low, which also suggests that effect sizes detected by models were small and that these models were underpowered [51, 52], making them more sensitive to variations.

In the original study, authors also reported performance of the models evaluated on an external dataset (Table 2 of [1]) and with Leave-One-Site-Out cross-validation (LOSO CV) in the outer-loop compared to Leave-One-Out (LOO CV) in the main study. They found similar performance at Year 1 (R2 over 0.5) with these variations compared to the main results. Performance at other time points was not available for the external validation, but for LOSO CV, models trained for prediction at Year 2 also performed very well and those of time point Baseline and Year 4 exhibited lower prediction ability

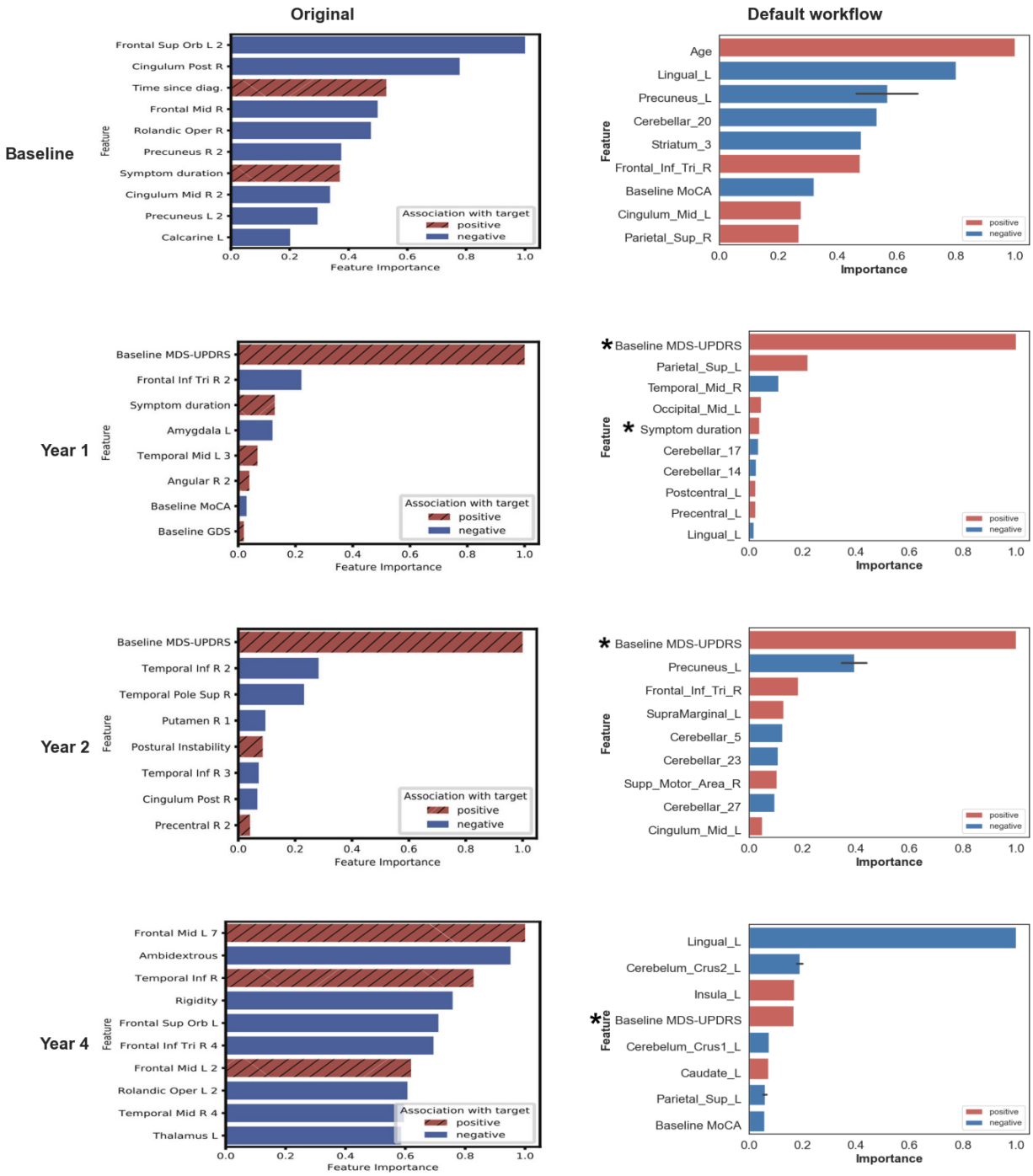

Figure 5. Predictive features learned by the best performing models to predict MDS-UPDRS score at each time point for the original study (left - extracted from [1]) and the *default workflow* (right) using ReHo. Features with low importance were not shown. Red bars indicate a positive association and blue bars indicate a negative association. Stars (*) represent the presence of this feature in the original study and the reproduction.

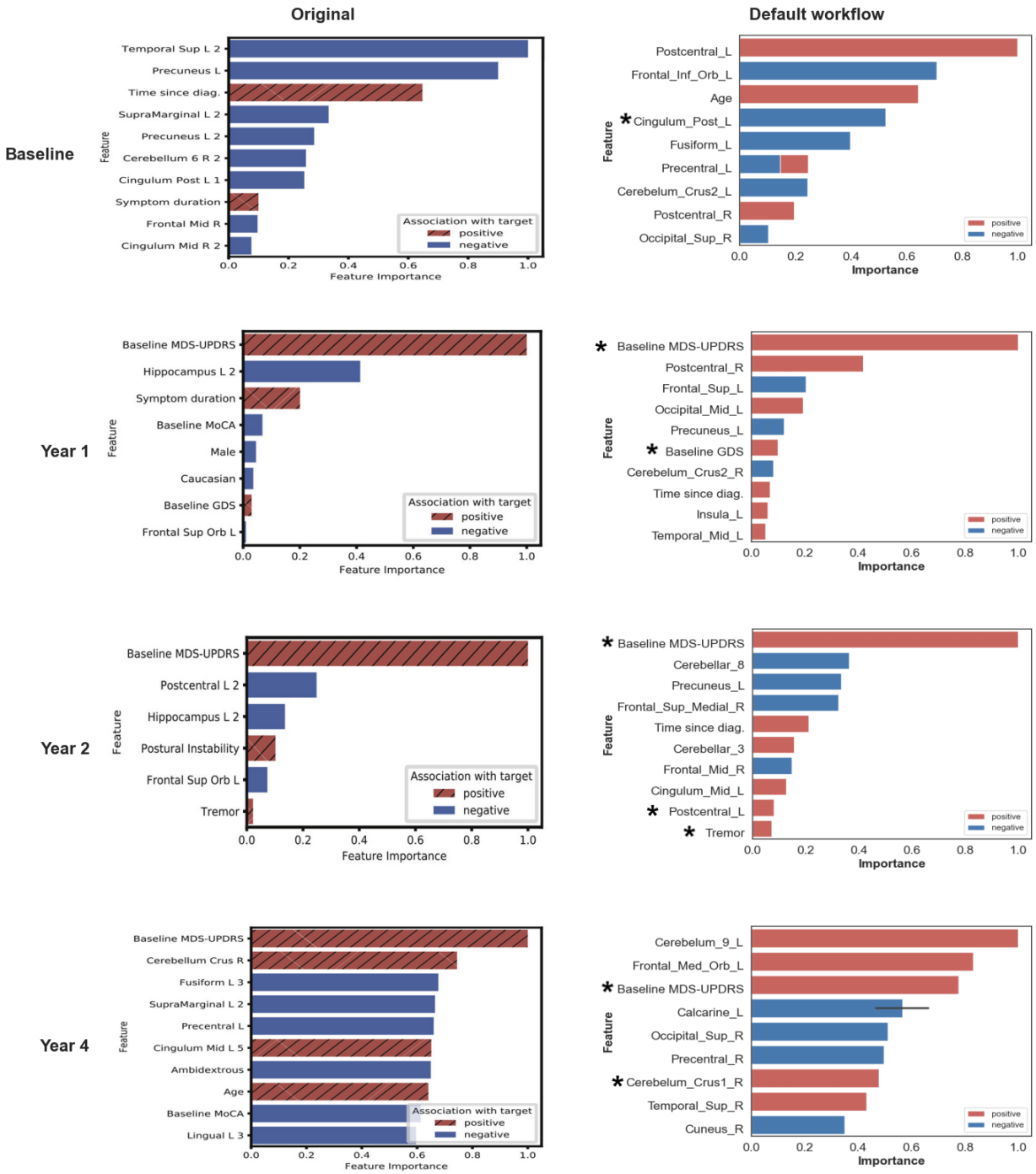

Figure 6. Predictive features learned by the best performing models to predict MDS-UPDRS score at each time point for the original study (left - extracted from [1]) and the *default workflow* (right) using fALFF. Features with low importance were not shown. Red bars indicate a positive association and blue bars indicate a negative association. Stars (*) represent the presence of this feature in the original study and the reproduction.

compared to the ones tuned using the LOO CV scheme (main original workflow). These two comparisons are consistent with our observations on the robustness to image features variations of models at Year 1 and Year 2 and the high sensitivity of models at Baseline and Year 4.

When using a different cohort with distinctions in the distribution of the most important feature (MDS-UPDRS score at Baseline) of the Year 1 model, a lower performance was found using fALFF ($p < 0.05$) and ReHo. This performance remained high and close to the one reported in the original study. Moreover, when removing specific clinical features such as MDS-UPDRS Baseline scores, the performance models at Year 1 and Year 2 significantly dropped. This suggests that the robustness mentioned above was probably dependant on the distribution of these measures. It would be interesting to assess the interaction of variations in both cohorts, imaging features and input features sets to see if the robustness to analytical variations was also present using the replication cohorts and when increasing the importance of image features in the prediction.

Challenges of reproducibility studies

In our reproducibility study, several challenges were encountered, in particular related to cohort selection, fMRI feature pre-processing, and results reporting. To extract the same Baseline cohort as used in [1], we first attempted to query the PPMI database using the information available in the paper and the code shared by the authors. This step was unsuccessful since we could not get the same sample size at Baseline (102 instead of 82 in [1]), and we decided to contact the authors who provided us the exact subject and visit list used in the original study. With this list, we were able to build a cohort with the same participants at Baseline. A potential solution to avoid similar difficulties in future reproducibility studies would be to register cohorts obtained from public databases under the same data usage agreements as the original data. In the case of PPMI, a specific section of the online portal could be created to store cohort definitions and associate them with published manuscripts.

Even with the original participant identifiers and visit list at Baseline, we could not retrieve the same Baseline cohort in the PPMI database. Our closest-to-original cohort included the 82 original participants, but for 5 of them, a different visit than the original one was used. For 3 of these visits, we intentionally chose to keep the visits selected by our first query to better fit with the description of the cohort in the paper. For the 2 other visits, the functional images corresponding to these participants and visits were not available anymore in the PPMI database. Since the PPMI database continuously adds new participant visits, we chose to keep only the visits that were added more than a year before the original study publication, since the original authors did not report the date at which they queried the database. With this filter, the Baseline participants list and the exact same code used to search for follow-up visits, the cohorts obtained for follow-up visits were still dissimilar to the original ones, with more participants and several noteworthy differences in clinical and demographic variables. A first step to solve this particular issue would be to systematically report the date when databases are queried. However, the issues faced when attempting to reproduce the original cohort in fact highlight the need for version control in public databases, using tools such as DataLad [53] that is for instance adopted in the OpenNeuro database [54]. With version control, authors would be able to cite the exact version of the database used, which would importantly facilitate cohort reproductions.

Reproducing the fMRI pre-processing and feature computation pipelines described in [1] also raised challenges. First, although authors provided a description of the different pre-processing steps performed and tools used, exact reproductions of neuroimaging pipelines require more detailed information — including specific parameters values,

name and version of the standard template used, software versions — given the overall complexity and flexibility of image analysis methods [55]. To reproduce the pipeline used in [1], we had to make informed guesses about important parameters of the analysis. Some of these choices were conditioned by the nature of the neuroimaging pipelines (e.g., the choice of standard template to register functional images was constrained by the use of ICA-AROMA) while other decisions were more arbitrary and led to multiple valid variations (e.g., the computation of WM and CSF mean time-series for which we applied three different variations with different software packages and methods). Reporting guidelines, such as COBIDAS [56], were developed to help document analyses and facilitate reproduction studies. However, to reproduce complete analyses, sharing the entirety of the code used in the original experiment remains the most valuable information, as it contains a both human and machine-readable description of the exact method employed. Code-sharing platforms such as GitHub and GitLab are now widely available for this purpose and long-term preservation of these code is supported by archive systems such as Software Heritage [57, 58] or Zenodo.

The use of a custom-based atlas to parcellate the brain in the original study also created challenges. Future reproducibility studies would benefit from comprehensive descriptions of the methods used to create such custom data, access to the code to create the data, and sharing of the data itself through platforms such as Zenodo, the Open-Science Framework, Figshare, or NeuroVault [59]. Such platforms could also be used for sharing derived data, for instance whole-brain fALFF and ReHo maps. However, Data Usage Agreements often requires that derived data have to be shared under the same conditions. We emphasize again the need for specific platforms in public databases to host data associated with a published manuscript, including cohort descriptions and derived imaging data.

The authors of [1] shared code used in the original study, in particular for feature computation (fALFF and ReHo after pre-processing and clinical/demographic features search in PPMI study files) and machine-learning models training. The availability of this code was extremely useful for our reproducibility study, and we warmly acknowledge the authors for taking the time to share reusable code with their analysis. Despite the availability of the code, we still faced some difficulties to reproduce the results presented in the original study, due to discrepancies between the methods reported in the paper and the code shared, especially for the imaging feature computation, the cross-validation procedure and the results reports. For instance, we were not able to retrieve the Z-scoring of whole-brain fALFF and ReHo maps mentioned in the paper. This discrepancy was likely due to the update of the C-PAC pipeline used by the authors for pre-processing, in which the documentation still mentioned the possibility to output Z-scored maps even if this option was not implemented anymore in the pipeline. This reiterate the importance of code versioning and reporting software versions. The use of software container engines such as Docker and Singularity in combination with frameworks such as Boutiques [37] or BIDS-Apps [60] facilitates reproduction and reduces the technical work required to find and install the software versions used in the original study.

Regarding model selection and optimization, we highlight the complexity of nested cross-validation schemes and the on-going debate on the choice of rigorous cross-validation procedures [61, 20]. Here again, code sharing is required to describe the exact evaluation method used in the original study. At this level in the analysis, Jupyter notebooks [62] are an interesting option to document code and mix it with data, natural text and figures. Initiatives were recently launched to share reproducible Jupyter notebooks, such as NeuroLibre [63], a platform for sharing re-executable preprints. We created a Jupyter notebook for our study, that we made publicly available at https://github.com/elodiegermani/nguyen-etal-2021.

To conclude, we highlight the challenges associated with the reproduction of neu-

roimaging studies. We discussed some of the specific difficulties encountered in our study, and provided some potential solutions to facilitate this process in the future, in terms of time cost and adequacy of the reproduction. Nevertheless, given the complexity of the data, software and analyses required in current neuroimaging studies, reproducing existing papers remains extremely challenging.

Acknowledgments

This work was funded by the Michael J. Fox Foundation for Parkinson's Research (MJFF-021134). This work was also funded by a MITACS Global Research Award (IT34055). This work was partially funded by Region Bretagne (ARED MAPIS) and Agence Nationale pour la Recherche for the programm of doctoral contracts in artificial intelligence (project ANR-20-THIA-0018).

References

1. Kevin P. Nguyen, Vyom Raval, Alex Treacher, Cooper Mellema, Fang Frank Yu, Marco C. Pinho, Rathan M. Subramaniam, Richard B. Dewey, and Albert A. Montillo. Predicting Parkinson's disease trajectory using clinical and neuroimaging baseline measures. *Parkinsonism & Related Disorders*, 85:44–51, 2021. ISSN 1353-8020. doi: 10.1016/j.parkreldis.2021.02.026.

2. Bastiaan R Bloem, Michael S Okun, and Christine Klein. Parkinson's disease. *The Lancet*, 397(10291):2284–2303, 2021. ISSN 0140-6736. doi: 10.1016/S0140-6736(21)00218-X.

3. Katrina Gwinn, Karen K David, Christine Swanson-Fischer, Roger Albin, Coryse St Hillaire-Clarke, Beth-Anne Sieber, Codrin Lungu, F DuBois Bowman, Roy N Alcalay, Debra Babcock, Ted M Dawson, Richard B Dewey, Tatiana Foroud, Dwight German, Xuemei Huang, Vlad Petyuk, Judith A Potashkin, Rachel Saunders-Pullman, Margaret Sutherland, David R Walt, Andrew B West, Jing Zhang, Alice Chen-Plotkin, Clemens R Scherzer, David E Vaillancourt, and Liana S Rosenthal. Parkinson's disease biomarkers: perspective from the NINDS Parkinson's Disease Biomarkers Program. *Biomarkers in Medicine*, 11(6):451–473, 2017. ISSN 1752-0363. doi: 10.2217/bmm-2016-0370.

4. Trina Mitchell, Stéphane Lehéricy, Shannon Y. Chiu, Antonio P. Strafella, A. Jon Stoessl, and David E. Vaillancourt. Emerging Neuroimaging Biomarkers Across Disease Stage in Parkinson Disease: A Review. *JAMA Neurology*, 78(10):1262–1272, 2021. ISSN 2168-6149. doi: 10.1001/jamaneurol.2021.1312. URL https://doi.org/10.1001/jamaneurol.2021.1312.

5. Yanbing Hou and Huifang Shang. Magnetic Resonance Imaging Markers for Cognitive Impairment in Parkinson's Disease: Current View. *Frontiers in Aging Neuroscience*, 14, 2022. ISSN 1663-4365. doi: 10.3389/fnagi.2022.788846.

6. Jason D. Warren, Jonathan D. Rohrer, Jonathan M. Schott, Nick C. Fox, John Hardy, and Martin N. Rossor. Molecular nexopathies: a new paradigm of neurodegenerative disease. *Trends in Neurosciences*, 36(10):561–569, 2013. ISSN 0166-2236. doi: 10.1016/j.tins.2013.06.007.

7. YanBing Hou, ChunYan Luo, Jing Yang, RuWei Ou, Wei Song, QianQian Wei, Bei Cao, Bi Zhao, Ying Wu, Hui-Fang Shang, and QiYong Gong. Prediction of

individual clinical scores in patients with Parkinson's disease using resting-state functional magnetic resonance imaging. *Journal of the Neurological Sciences*, 366: 27–32, 2016. ISSN 0022-510X. doi: 10.1016/j.jns.2016.04.030.

8. Xiao-Fei Hu, Jiu-Quan Zhang, Xiao-Mei Jiang, Chao-Yang Zhou, Lu-Qing Wei, Xun-Tao Yin, Jing Li, Yan-Ling Zhang, and Jian Wang. Amplitude of Low-frequency Oscillations in Parkinson's Disease: A 2-year Longitudinal Resting-state Functional Magnetic Resonance Imaging Study. *Chinese Medical Journal*, 128 (05):593–601, 2015. doi: 10.4103/0366-6999.151652.

9. HuiZe Pang, ZiYang Yu, HongMei Yu, JiBin Cao, YingMei Li, MiaoRan Guo, ChengHao Cao, and GuoGuang Fan. Use of machine learning method on automatic classification of motor subtype of Parkinson's disease based on multilevel indices of rs-fMRI. *Parkinsonism & Related Disorders*, 90:65–72, 2021. ISSN 1353-8020. doi: 10.1016/j.parkreldis.2021.08.003.

10. Tao Wu, Xiangyu Long, Yufeng Zang, Liang Wang, Mark Hallett, Kuncheng Li, and Piu Chan. Regional homogeneity changes in patients with Parkinson's disease. *Human Brain Mapping*, 30(5):1502–1510, 2009. ISSN 1097-0193. doi: 10.1002/hbm.20622.

11. Yumei Yue, Yasi Jiang, Ting Shen, Jiali Pu, Hsin-Yi Lai, and Baorong Zhang. ALFF and ReHo Mapping Reveals Different Functional Patterns in Early- and Late-Onset Parkinson's Disease. *Frontiers in Neuroscience*, 14, 2020. ISSN 1662-453X. doi: doi.org/10.3389/fnins.2020.00141.

12. Yufeng Zang, Tianzi Jiang, Yingli Lu, Yong He, and Lixia Tian. Regional homogeneity approach to fMRI data analysis. *NeuroImage*, 22(1):394–400, 2004. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2003.12.030.

13. Qi-Hong Zou, Chao-Zhe Zhu, Yihong Yang, Xi-Nian Zuo, Xiang-Yu Long, Qing-Jiu Cao, Yu-Feng Wang, and Yu-Feng Zang. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *Journal of Neuroscience Methods*, 172(1):137–141, 2008. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2008.04.012.

14. Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366, 2011. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797611417632.

15. Alexander Bowring, Camille Maumet, and Thomas E. Nichols. Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping*, 40(11): 3362–3384, 2019. ISSN 1065-9471, 1097-0193. doi: 10.1002/hbm.24603. Number: 11.

16. Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, Paolo Avesani, Blazej M Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, and Marco Barilari. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020. doi: https://doi.org/10.1038/s41586-020-2314-9.

17. Tristan Glatard, Lindsay B. Lewis, Rafael Ferreira da Silva, Reza Adalat, Natacha Beck, Claude Lepage, Pierre Rioux, Marc-Etienne Rousseau, Tarek Sherif, Ewa Deelman, Najmeh Khalili-Mahani, and Alan C. Evans. Reproducibility of

neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, 9, 2015. ISSN 1662-5196. doi: 10.3389/fninf.2015.00012.

18. Ed HBM Gronenschild, Petra Habets, Heidi IL Jacobs, Ron Mengelers, Nico Rozendaal, Jim Van Os, and Machteld Marcelis. The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, 7(6):e38234, 2012.

19. Sayash Kapoor and Arvind Narayanan. Leakage and the Reproducibility Crisis in ML-based Science, 2022.

20. Gael Varoquaux and Olivier Colliot. Evaluating Machine Learning Models and Their Diagnostic Value. In *Machine Learning for Brain Disorders*, Neuromethods, pages 601–630. Springer US, 2023. ISBN 978-1-07-163195-9. doi: 10.1007/978-1-0716-3195-9_20. URL https://doi.org/10.1007/978-1-0716-3195-9_20.

21. Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101694. URL https://www.sciencedirect.com/science/article/pii/S1361841520300591.

22. Lorena A. Barba. Terminologies for reproducible research, 2018.

23. Ullrich Wüllner, Per Borghammer, Chi-un Choe, Ilona Csoti, Björn Falkenburger, Thomas Gasser, Paul Lingor, and Peter Riederer. The heterogeneity of Parkinson's disease. *Journal of Neural Transmission*, 130(6):827–838, 2023. ISSN 0300-9564. doi: 10.1007/s00702-023-02635-4.

24. Simon Klau, Chirag J Patel, John PA Ioannidis, Anne-Laure Boulesteix, Sabine Hoffmann, et al. Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. *Meta-Psychology*, 7, 2023.

25. Russell A. Poldrack, Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126, 2017. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn.2016.167. URL http://www.nature.com/articles/nrn.2016.167. Number: 2.

26. Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M Tanner, John Q Trojanowski, Leslie M Shaw, John Seibyl, Norbert Schuff, Andrew Singleton, Karl Kieburtz, Arthur W Toga, Brit Mollenhauer, Doug Galasko, Lana M Chahine, Daniel Weintraub, Tatiana Foroud, Duygu Tosun-Turgut, Kathleen Poston, Vanessa Arnedo, Mark Frasier, Todd Sherer, and Parkinson's Progression Markers Initiative. The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477, 2018. ISSN 2328-9503. doi: 10.1002/acn3.644.

27. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman,

Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

28. Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3(1):160044, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.44.

29. Yaroslav Halchenko, Mathias Goncalves, Pablo Velasco, Visconti di Oleggio Castello Matteo, and Satrajit Ghosh. nipy/heudiconv: v0.13.1. doi: 10.5281/zenodo.7963413. https://zenodo.org/records/7963413.

30. Mark Jenkinson, Christian F. Beckmann, Timothy E. J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *NeuroImage*, 62(2):782–790, 2012. ISSN 1095-9572(Electronic),1053-8119(Print). doi: 10.1016/j.neuroimage.2011.09.015.

31. Robert W. Cox. AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, 29(3): 162–173, 1996. ISSN 0010-4809. doi: 10.1006/cbmr.1996.0014.

32. Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.09.025.

33. Raimon H.R. Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, 112:267–277, 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.02.064.

34. Krzysztof Gorgolewski. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, page 15, 2017. doi: 10.5281/zenodo.581704.

35. Neurodocker. URL https://github.com/ReproNim/neurodocker.

36. Anaconda software distribution. *Anaconda Documentation*, 2020. https://docs.anaconda.com/.

37. Tristan Glatard, Gregory Kiar, Tristan Aumentado-Armstrong, Natacha Beck, Pierre Bellec, Rémi Bernard, Axel Bonnet, Sorina Camarasu-Pop, Frédéric Cervenansky, Samir Das, Rafael Ferreira da Silva, Guillaume Flandin, Pascal Girard, Krzysztof J. Gorgolewski, Charles R. G. Guttmann, Valérie Hayot-Sasson, Pierre-Olivier Quirion, Pierre Rioux, Marc-Eienne Rousseau, and Alan C. Evans. Boutiques: a flexible framework for automated application integration in computing platforms, 2017.

38. Elodie Germani. Image processing, 2023. URL https://doi.org/10.5281/zenodo.10298335.

39. Elodie Germani. Trainer, 2023. URL https://doi.org/10.5281/zenodo.10298359.

40. ICA-AROMA & fmriprep using child template - fmriprep - Neurostars. URL https://neurostars.org/t/ica-aroma-fmriprep-using-child-template/5139.

41. Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00014.

42. Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019. ISSN 1548-7105. doi: 10.1038/s41592-018-0235-4.

43. Jonathan D. Power, Anish Mitra, Timothy O. Laumann, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84:320–341, 2014. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2013.08.048. URL https://www.sciencedirect.com/science/article/pii/S1053811913009117.

44. Linden Parkes, Ben Fulcher, Murat Yücel, and Alex Fornito. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, 171:415–436, 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2017.12.073. URL https://www.sciencedirect.com/science/article/pii/S1053811917310972.

45. Craddock Cameron, Sikka Sharad, Cheung Brian, Khanuja Ranjeet, Ghosh Satrajit, Yan Chaogan, Li Qingyang, Lurie Daniel, Vogelstein Joshua, Burns Randal, Colcombe Stanley, Mennes Maarten, Kelly Clare, Di Martino Adriana, Castellanos Francisco, and Milham Michael. Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). *Frontiers in Neuroinformatics*, 7, 2013. ISSN 1662-5196. doi: 10.3389/conf.fninf.2013.09.00042.

46. Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and B T Thomas Yeo. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, September 2018. ISSN 1047-3211. doi: 10.1093/cercor/bhx179.

47. Pierre Bellec, Pedro Rosa-Neto, Oliver C. Lyttelton, Habib Benali, and Alan C. Evans. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, 51(3):1126–1139, 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.02.082.

48. Jörn Diedrichsen, Joshua H. Balsters, Jonathan Flavell, Emma Cussans, and Narender Ramnani. A probabilistic MR atlas of the human cerebellum. *NeuroImage*, 46(1):39–46, 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2009.01.045.

49. Andri C. Tziortzi, Suzanne N. Haber, Graham E. Searle, Charalampos Tsoumpas, Christopher J. Long, Paul Shotbolt, Gwenaelle Douaud, Saad Jbabdi, Timothy E. J. Behrens, Eugenii A. Rabiner, Mark Jenkinson, and Roger N. Gunn. Connectivity-Based Functional Analysis of Dopamine Release in the Striatum Using Diffusion-Weighted MRI and Positron Emission Tomography. *Cerebral Cortex*, 24(5): 1165–1177, 2014. ISSN 1047-3211. doi: 10.1093/cercor/bhs397.

50. Edmund T. Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, and Marc Joliot. Automated anatomical labelling atlas 3. *NeuroImage*, 206:116189, 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116189. URL https://www.sciencedirect.com/science/article/pii/S1053811919307803.

51. Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn3475. URL http://www.nature.com/articles/nrn3475.

52. John P. A. Ioannidis. Why Most Discovered True Associations Are Inflated:. *Epidemiology*, 19(5):640–648, 2008. ISSN 1044-3983. doi: 10.1097/EDE.0b013e31818131e7. URL http://journals.lww.com/00001648-200809000-00002.

53. Yaroslav O. Halchenko, Kyle Meyer, Benjamin Poldrack, Debanjum Singh Solanky, Adina S. Wagner, Jason Gors, Dave MacFarlane, Dorian Pustina, Vanessa Sochat, Satrajit S. Ghosh, Christian Mönch, Christopher J. Markiewicz, Laura Waite, Ilya Shlyakhter, Alejandro de la Vega, Soichi Hayashi, Christian Olaf Häusler, Jean-Baptiste Poline, Tobias Kadelka, Kusti Skytén, Dorota Jarecka, David Kennedy, Ted Strauss, Matt Cieslak, Peter Vavra, Horea-Ioan Ioanas, Robin Schneider, Mika Pflüger, James V. Haxby, Simon B. Eickhoff, and Michael Hanke. DataLad: distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software*, 6(63):3262, 2021. ISSN 2475-9066. doi: 10.21105/joss.03262. URL https://joss.theoj.org/papers/10.21105/joss.03262.

54. Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, Anita Jwa, and Russell Poldrack. The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10:e71774, 2021. ISSN 2050-084X. doi: 10.7554/eLife.71774. URL https://doi.org/10.7554/eLife.71774.

55. Joshua Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, page 13, 2012.

56. Thomas E Nichols, Samir Das, Simon B Eickhoff, Alan C Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P Milham, Russell A Poldrack, Jean-Baptiste Poline, Erika Proal, Bertrand Thirion, David C Van Essen, Tonya White, and B T Thomas Yeo. Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, 20(3):299–303, 2017. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4500. URL http://www.nature.com/articles/nn.4500.

57. Roberto Di Cosmo and Stefano Zacchiroli. Software heritage: Why and how to preserve software source code. In *iPRES 2017: 14th International Conference on Digital Preservation*, 2017. URL https://www.softwareheritage.org/wp-content/uploads/2020/01/ipres-2017-swh.pdfhttps://hal.archives-ouvertes.fr/hal-01590958.

58. Jean-François Abramatic, Roberto Di Cosmo, and Stefano Zacchiroli. Building the universal archive of source code. *Communications of the ACM*, 61(10):29–31, 2018. ISSN 0001-0782. doi: 10.1145/3183558. URL https://cacm.acm.org/magazines/2018/10/231366-building-the-universal-archive-of-source-code/fulltext.

59. Krzysztof J. Gorgolewski, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S. Ghosh, Camille Maumet, Vanessa V. Sochat, Thomas E. Nichols, Russell A. Poldrack, Jean-Baptiste Poline, Tal Yarkoni, and Daniel S. Margulies. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9, 2015. ISSN 1662-5196. doi: 10.3389/fninf.2015.00008. URL https://www.frontiersin.org/article/10.3389/fninf.2015.00008.

60. Krzysztof J Gorgolewski, Fidel Alfaro-Almagro, Tibor Auer, Pierre Bellec, Mihai Capotă, M Mallar Chakravarty, Nathan W Churchill, Alexander Li Cohen, R Cameron Craddock, Gabriel A Devenyi, et al. Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS computational biology*, 13(3):e1005209, 2017.

61. Jacques Wainer and Gavin Cawley. Nested cross-validation when selecting classifiers is overzealous for most practical applications, 2018. URL http://arxiv.org/abs/1809.09446.

62. Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

63. Elizabeth DuPre, Chris Holdgraf, Agah Karakuzu, Loïc Tetrel, Pierre Bellec, Nikola Stikov, and Jean-Baptiste Poline. Beyond advertising: New infrastructures for publishing integrated research objects. *PLOS Computational Biology*, 18(1): e1009651, 2022. doi: 10.1371/journal.pcbi.1009651. URL https://doi.org/10.1371/journal.pcbi.1009651.