Preprocessing, Prediction and Significance: Framework and Application to Brain Imaging

Martin Nørgaard[1,2], Brice Ozenne[1,4], Claus Svarer[1,2], Vibe G. Frokjaer[1], Martin Schain[1], Stephen C. Strother[5], and Melanie Ganz[1,3](✉)

[1] Neurobiology Research Unit, Rigshospitalet, Copenhagen, Denmark
mganz@nru.dk
[2] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[3] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[4] Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark
[5] Rotman Research Institute, Baycrest, University of Toronto, Toronto, Canada

Abstract. Brain imaging studies have set the stage for measuring brain function in psychiatric disorders, such as depression, with the goal of developing effective treatment strategies. However, data arising from such studies are often hampered by noise confounds such as motion-related artifacts, affecting both the spatial and temporal correlation structure of the data. Failure to adequately control for these types of noise can have significant impact on subsequent statistical analyses. In this paper, we demonstrate a framework for extending the non-parametric testing of statistical significance in predictive modeling by including a plausible set of preprocessing strategies to measure the predictive power. Our approach adopts permutation tests to estimate how likely we are to obtain a given predictive performance in an independent sample, depending on the preprocessing strategy used to generate the data. We demonstrate and apply the framework on examples of longitudinal Positron Emission Tomography (PET) data following a pharmacological intervention.

1 Introduction

Modern neuroimaging studies are complicated and comprised of many steps including subject selection, data acquisition, preprocessing and some form of statistical analysis. In the past decade, there has been a growing concern about the validity and reproducibility of produced findings in such studies, and this has largely been attributed to low statistical power, software errors and flexible data analysis strategies [1,10].

Data sharing initiatives such as OpenNeuro (openneuro.org) are now enabling researchers to open up the subject selection and data acquisition factors of a

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11767, pp. 196–204, 2019.
https://doi.org/10.1007/978-3-030-32251-9_22

study by sharing raw image data publicly. Statistical analysis tools are also widely available in the major neuroimaging software packages (e.g. SPM, FSL, AFNI and FreeSurfer) or on GitHub, and the outputs of statistical analyses can be shared (e.g. on Neurovault). Furthermore, the analysis and statistical methods have been under intense scrutiny in the last years and concerns about errors in software packages as well as in the appropriate application of statistical methods have been heatedly discussed [2,4].

Conversely, the influence of preprocessing on the outcome of the data analysis has besides a few initiatives in fMRI [2,3] been an overlooked factor. Many laboratories have set up preprocessing pipelines that are used for all their studies and large research collaborations such as the Human Brain Project (HBP) have implemented a single preprocessing pipeline[1] that is used daily to extract features from subjects enrolled in neuroscience research studies. Hence, while researchers are focusing intensely on new statistical model development, the interaction of different types of preprocessing steps with the following statistical analysis is largely ignored [3].

One solution to limit the "researcher degrees of freedom" that has been proposed is the pre-registration of complete analysis pipelines e.g. with the Open Science Framework or AsPredicted [10]. The argument for pre-registration is that researchers should not be constrained to a single analysis method, but rather predefine which approach they will use. Additionally, there might not even exist a single best workflow for all studies of a given type, even though there is evidence that different workflows might be optimal for different studies or even for different individuals [3]. However, at the same time it seems to be implausible that out of thousands of possible workflows only the chosen pre-registered one would be able to show a true biological effect. It is much more likely that a range of different processing pipelines would have yielded the same conclusion of a given study. In the case of a strong effect, one might even hope that most processing pipelines - so no matter how the data has been preprocessed - would be able to detect the effect. Hence, it is also of interest to analyze not only the variance arising from the preprocessing [2,3], but to take the step further and analyze the variance that different preprocessing pipelines add to the statistical analysis of a study and its conclusions. On the one hand, this approach can highlight spurious findings due to a specific preprocessing pipeline, since most preprocessing pipelines would not be able to produce the same result. On the other, it can also give strong evidence for an effect if most preprocessing pipelines arrive at the same or very similar result.

In this work, we present a comprehensive framework to test the influence of preprocessing choices on the subsequent statistical analysis. We demonstrate how the choice of preprocessing can affect our belief in the available sample data, $\mathbf{x}$, with class labels y, to generalize to the true underlying joint distribution $p(\mathbf{x}, y)$. Our approach adopts a range of preprocessing choices as a generative model for $\mathbf{x}$, and evaluates the predictive performance for the conditional distribution $p(y|\mathbf{x})$ using permutations [6] and the maximum statistic [8]. By permuting and

[1] See https://github.com/HBPMedical/mri-preprocessing-pipeline.

evaluating across preprocessing choices, the framework provides a measure of how likely we are to obtain the observed prediction by chance, only because the preprocessing strategy interacted with the predictive model to identify a pattern that happened to correlate with the class labels. We first detail the framework and then give an example of its application based on a published study involving the serotonin transporter and PET imaging [5].

2 Non-parametric Framework for Joining Multiple Preprocessing Strategies with Prediction

The framework that we are proposing can roughly be broken into three major components: **(A)** definition of a subset of equally plausible preprocessing strategies, **(B1)** definition of the set of predictive models and the performance metric, **(B2)** cross validation to select the optimal predictive model and estimate the prediction, and **(C)** estimation of the statistical significance of the prediction accuracy (Fig. 1).

2.1 Defining a Subset of Preprocessing Strategies

In all fields of neuroimaging, before any statistical model is applied to a given data set $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, with N observations, the data is commonly preprocessed using a set of steps such as motion correction, co-registration and partial volume correction (Fig. 1A). The data set $\mathbf{x}_n \in \mathbf{R}^p$ are observations with p features and $y_n \in \{-1, 1\}$ are the corresponding class labels. The entire sequence of preprocessing steps is often referred to as a pipeline, designed to remove artifacts and noise from the data. Designing the optimal sequence of preprocessing steps is a challenging problem, mainly due to the high dimensionality of the data and due to the complex spatio-temporal noise structure. Therefore, several preprocessing algorithms have been proposed and refined over the years, with limited consensus in the community on the optimal strategy. The preprocessed data can for pipeline j be defined as $\{(\mathbf{x}_{n,j}, y_n)\}_{n=1}^N$.

2.2 Model Selection and Cross-Validation

Once the data has been preprocessed it is ready for statistical analysis. Next, we need to (1) select a predictive model and tune the model parameters to the data, and (2) assess the chosen predictive model by estimating its ability to predict on unseen data. For both (1) and (2), one common approach is to use cross-validation and evaluate the model in an independent test set (Fig. 1B). For this purpose, the data has to be randomly divided into a training data set and validation set. The training data may be further split into an inner cross-validation loop (nested cross-validation) [11]. Finally, the entire cross-validation has to be repeated M times to obtain an unbiased mean predictive accuracy. This approach aligns with community guidelines on model selection and cross-validation [11].

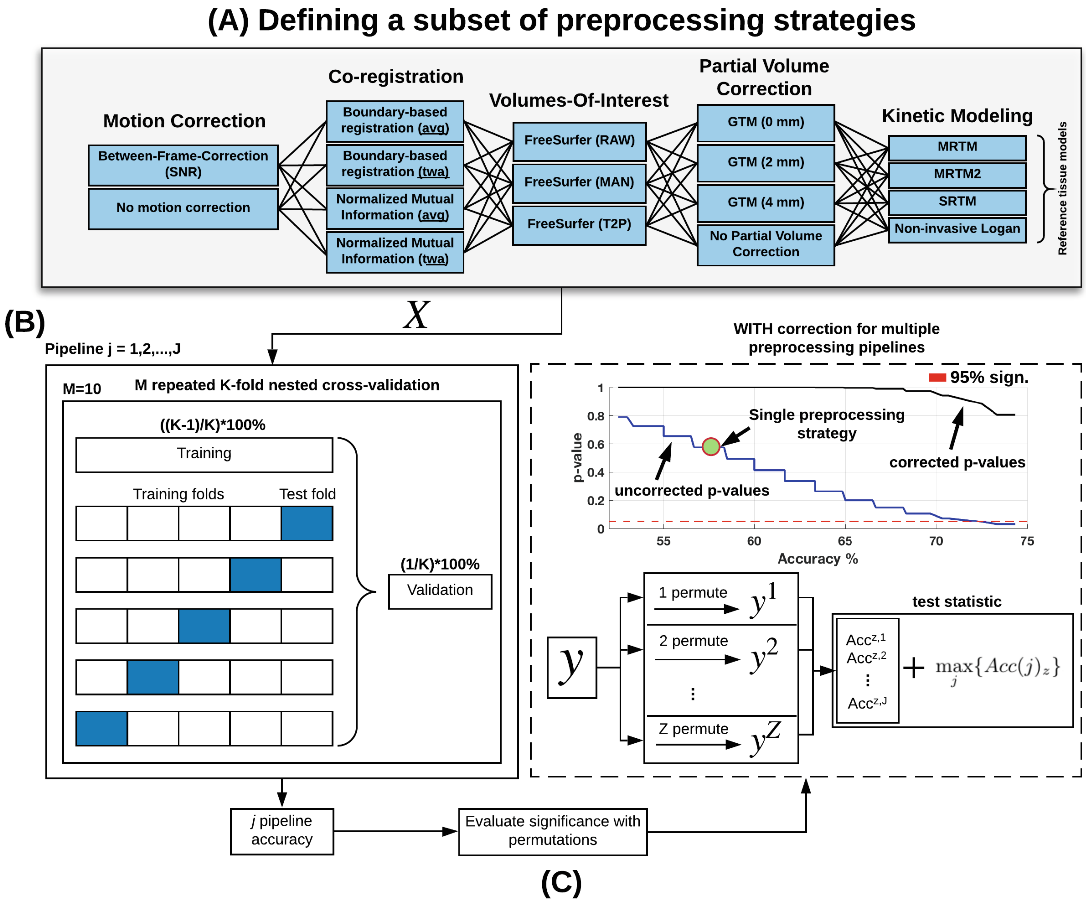

Fig. 1. (A) Definition of a subset of preprocessing strategies $j = 1, ..., J$: This includes preprocessing steps such as motion correction, co-registration, delineation of volumes of interest, partial volume correction, and kinetic modeling. **(B)** Model selection and cross-validation: For each pipeline j, select a classification model (e.g. Linear Discriminant), and a K-fold nested cross-validation scheme with M repetitions. **(C)** Evaluate the significance with permutations: Randomly permute the class labels $y \in \{-1, 1\}$, and re-run (B) for each pipeline j to obtain a classification accuracy for the $z = 1, .., Z$ permutation. For each permutation z select the maximum accuracy across all preprocessing pipelines J and for Z permutations generate a null-distribution of maximum accuracies. Use the null-distribution of the max-accuracies to obtain the p-value for each pipeline at a significance level α. NOTE: uncorrected p-values refer to original accuracies according to their permuted null-distribution at a significance level α.

2.3 Permutation Test for a Single Pipeline

Once a model has been selected and evaluated to provide a predictive accuracy, the gold standard is to estimate the statistical significance of the observed accuracy using permutations (Fig. 1C). The significance of each model and pipeline is estimated by randomly permuting the class labels Z times (i.e., sampling a permutation z from a uniform distribution π^z over the set, $\mathbf{\Pi}_N$, of all permutations of indices $1, ..., N$) and re-running the above M times repeated K-fold cross-validation procedure, and after Z replications generate an empirical null-distribution. This distribution may be used to obtain an empirical p-value for each model at an acceptable significance level α. Normally, this would be the last step of the data analysis. However, even though nested cross-validation can tune model parameters while avoiding circularity bias, there is still a hidden multiple comparison problem following the application of different preprocessing strategies. We therefore propose an extension to the current guidelines, by introducing a test statistic of maximum accuracies across equally plausible preprocessing pipelines. This approach should have a strong control over experiment-wise type I error.

2.4 Permutation Test for Multiple Pipelines

Rather than computing the permutation distribution of the accuracy for a single preprocessing pipeline j, we compute the permutation distribution of the maximum accuracy across all preprocessing pipelines. Let $\mathbf{\Pi}_N$ be a set of all permutations of indices $1, ..., N$, where N is the number of independent observations in the data set. The permutation test procedure that consists of Z iterations is defined as follows:

- Repeat Z times (with index $z = 1, ..., Z$)
  - sample a permutation π^z from a uniform distribution over $\mathbf{\Pi}_N$
  - compute the accuracy for each pipeline j for this permutation of labels
  - save the maximum accuracy across all pipelines J

$$t^z_{max} = \max_j \{Acc(\mathbf{x}_{1,j}, y_{\pi^z_1}, ..., \mathbf{x}_{N,j}, y_{\pi^z_N})\}$$

- Construct an empirical cumulative distribution of maximum accuracies

$$\hat{P}_{max}(T \leq t) = \frac{1}{Z} \sum_{1=z}^{Z} \Theta(t - t^z_{max})$$

  where Θ is a Heaviside step function ($\Theta(x) = 1$, if $x \geq 0$; 0 otherwise).
- Compute the accuracy for the true labels (non-permuted) for each pipeline j, $t_{0,j} = Acc(\mathbf{x}_{1,j}, y_1, ..., \mathbf{x}_{N,j}, y_N)$, and its corresponding p-value p^j_0 under the empirical distribution $\hat{P}_{max}$.

In our case, the null hypothesis assumes that the two classes have identical distributions, $\forall \mathbf{x} : p(\mathbf{x}|y = 1) = p(\mathbf{x}|y = -1)$, hence we deal with class balanced data. We reject the null hypothesis at level α if the accuracy for the true labeling of the data is in the α times 100% of the permuted distribution of the maximum accuracy. We can reject the null hypothesis for any preprocessing pipeline with an accuracy exceeding this threshold.

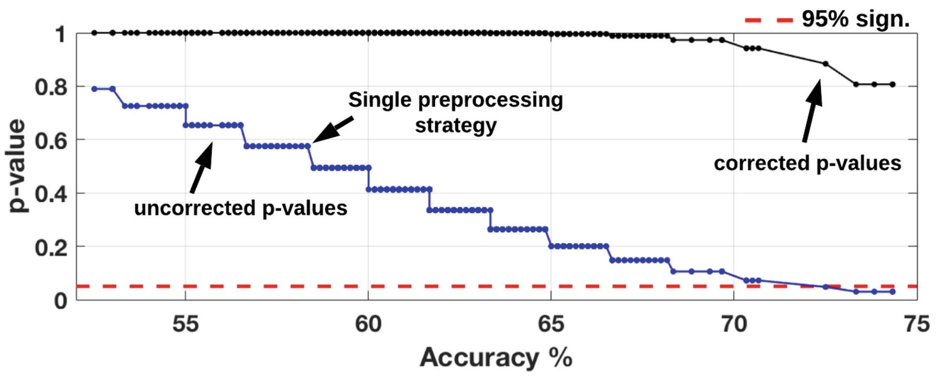

Fig. 2. Accuracy as a function of p-value for 384 preprocessing strategies. The blue dots indicate the p-value for each pipeline according to its permuted null distribution (uncorrected) and the black dots indicate the p-value according to the maximum permuted null distribution (corrected). The red dotted line is the 95% significance level.

2.5 Use of the Maximum Statistic in Neuroimaging

Correction of p-values using the maximum statistic has been used before in statistical studies of neuroimaging data [7,8]. Furthermore, several studies have examined the effects of multiple preprocessing options in combination with prediction [2,3]. The latter studies mainly focused on increasing predictive accuracy by examining multiple preprocessing strategies, but did not evaluate the prediction relative to random. Our work extends the non-parametric testing of statistical significance in predictive modeling by including a plausible set of preprocessing strategies to measure the predictive power.

3 Experiments

We illustrate the use of the framework in an experiment with data from a longitudinal PET study with a baseline and a re-scan, following a pharmacological intervention [5]. This data is publicly available through the CIMBI database (www.cimbi.dk). The data, $\mathbf{x}$, consists of 60 observations (30 baseline and 30 intervention scans) each with levels of serotonin transporter binding ($\mathrm{BP}_{\mathrm{ND}}$) in 34 cortical brain regions covering the entire neocortex. The corresponding class labels are $y_n \in \{baseline, intervention\}$. For quantification of $\mathrm{BP}_{\mathrm{ND}}$, we preprocessed the data using a fixed sequence of five preprocessing steps, each with varying parameter choices: (1) motion correction (with/without), (2) co-registration (four choices), (3) delineation of volumes-of-interest (three choices), (4) partial volume correction (four choices), and (5) kinetic modeling for quantification of $\mathrm{BP}_{\mathrm{ND}}$ (MRTM, SRTM, Non-invasive Logan and MRTM2). This results in $2 \times 3 \times 4^3 = 384$ combinations of preprocessing (Fig. 1A). Details are described in [9]. In the experiment, we used a Linear Discriminant classifier to predict the classes (baseline and intervention). The number of M repeated cross-validation iterations was 10, the number of K cross-validation folds was 5, and the number of permutation iterations Z was 1,000. To obtain true independence between

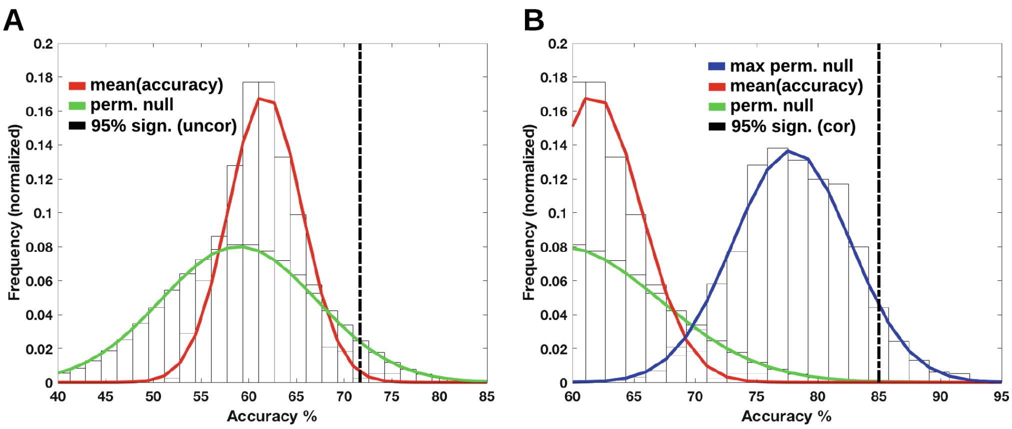

Fig. 3. **(A)** Average classification accuracies across preprocessing pipelines obtained using nested cross-validation with 10 repeats (red). The permuted null distribution of classification accuracies (1000 permutations) across preprocessing pipelines is visualized by the green distribution. The vertical dotted line is the 95% significance level of the permuted null distribution of classification accuracies across pipelines **(B)** The blue distribution is the permuted null distribution (1000 permutations) of maximum classification accuracies across preprocessing pipelines. The vertical dotted line is the 95% significance level for the permuted null distribution of maximum accuracies. (Color figure online)

the data and the labels, observations for each subject (i.e. baseline and intervention) were always stratified in the cross-validation. To summarize, the goal is to predict whether an observation in $\mathbf{x}$ is either a baseline or an intervention scan.

We start by studying the behaviour of accuracies and p-values, when varying the preprocessing strategy, reported in Fig. 2. Every point in Fig. 2 represents a preprocessing strategy with an accuracy and a p-value, either uncorrected (blue) or corrected (black). By changing the preprocessing strategy, this substantially improves the accuracy, with values ranging from 52% to 75%. There also exists a subset of preprocessing strategies that are significantly different ($p < 0.05$) from their permuted null distribution. The black line in Fig. 2 shows the p-values relative to the maximum permuted null distribution. The p-values decrease with increasing accuracy, but a much higher accuracy is needed compared to the blue line to obtain a significant p-value.

Figure 3 shows the distribution of accuracies for the estimated mean accuracies with the true labels (red), for the randomly permuted (green), and the maximum permuted (blue). Most preprocessing strategies fall within the permuted null distribution, but a subset of the preprocessing strategies are able to obtain statistical significance at $p < 0.05$ (i.e. less than 5% chance of observing better than 75% accuracy if the data and labels are truly independent). But to reject the null hypothesis under the empirical distribution of the maximum classification accuracies across pipelines, one would need an expected classification accuracy of 85% to obtain statistical significance at $\alpha = 0.05$ (Fig. 3).

4 Discussion and Conclusion

In this work, we extend the non-parametric testing of statistical significance in predictive modeling by including a plausible set of preprocessing strategies to measure the predictive power. We demonstrate its application in a longitudinal PET study. In this case, there are a few choices of preprocessing that lead to a significant prediction while the majority of preprocessing choices lead to a non-significant prediction (uncorrected). When correcting using knowledge about all the applied pipelines, no significant predictions survive (corrected using the maximum statistic).

While the statistical analysis of each individual preprocessing pipeline is carried out in an optimal fashion due to the use of M times repeated K-fold nested cross-validation, some of the preprocessing pipelines can still result in a significant prediction by chance. The reason for this can be that the preprocessing pipeline introduces spurious relations between the features and the labels, consequently overestimating the generalizability of the learning method. Our approach enables the examination of predictions across multiple preprocessing choices, providing a measure of the variance of the predictions across pipelines. Based on this we advise that care must be taken in a statistical analysis to avoid attributing an effect to a treatment/condition that was due to a single pipeline and/or predictive model.

The framework that we are proposing is not without limitations. First, while the goal of preprocessing is to factor out correlated features from the feature one is interested in, this is not necessarily a guarantee. For example, if one preprocessing strategy fails at factoring out correlated features and produces a “significant finding”, and a different pipeline correctly removes correlated features and produces a “non-significant finding”, this cannot be detected. This is one of the major drawbacks of data-driven selection of preprocessing strategies, and one risks drawing wrong conclusions if blindly selecting the preprocessing in a data-driven manner. In addition, as we are assuming independence between the preprocessing choices, one could worry that the effect we are observing, is simply due to the effect of assigning too much probability mass to strategies that no one would ever use. However, if we assume that all the included strategies are equally likely to be used, the proposed framework provides the researcher with a strong belief in the prediction under a set of plausible preprocessing strategies. This belief is both useful for the researcher carrying out the study, but also for colleagues reviewing the work for publication, as the impact of minor variations in acquisition/preprocessing is challenging to evaluate. The framework is also very flexible, and may be expanded to include a larger subset of preprocessing pipelines, a larger subset of features, but also a larger subset of statistical models (SVM, ANOVA, t-test etc.) with varying model complexities. However, it is noteworthy that the inclusion of more pipelines will also broaden the permuted null distribution further due to increased noise, so an increase in the number of pipelines will punish the ability to obtain statistical significance for any pipeline. The main point we hope to convey is that in future studies, researchers should not only pre-register their preprocessing or analysis as proposed by [10], but should also provide the variance of their results across a set of plausible preprocessing pipelines by using our framework. Because data acquisition is the most costly part of any experiment, spending resources on computing power by employing a framework as we propose is negligible in comparison. For future work, we still need to find a way of assigning appropriate non-uniform probability mass to strategies with different levels of relevance, otherwise we risk that the variance of the null distribution of maximum accuracies will be grossly overestimated. However, this is beyond the scope of this paper, and is left for future work.

References

1. Button, K.S., et al.: Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. **14**(5), 365 (2013)
2. Carp, J.: On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. Front. Neurosci. **6**, 149 (2012)
3. Churchill, N.W., et al.: An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. PLoS ONE **10**(7), e0131520 (2015)
4. Eklund, A., et al.: Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. PNAS **113**(28), 7900–7905 (2016)
5. Frokjaer, V.G., et al.: Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: a positron emission tomography study. Biol. Psychiatry **78**(8), 534–543 (2015)
6. Golland, P., Fischl, B.: Permutation tests for classification: towards statistical significance in image-based studies. In: Taylor, C., Noble, J.A. (eds.) IPMI 2003. LNCS, vol. 2732, pp. 330–341. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45087-0_28
7. Holmes, A.P., et al.: Nonparametric analysis of statistic images from functional mapping experiments. JCBFM **16**(1), 7–22 (1996)
8. Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. **15**(1), 1–25 (2002)
9. Nørgaard, M., et al.: Optimization of preprocessing strategies in Positron Emission Tomography (PET) neuroimaging: a [11C] DASB study. NeuroImage **199**, 466–479 (2019)
10. Poldrack, R.A., et al.: Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. **18**(2), 115 (2017)
11. Varoquaux, G., et al.: Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage **145**, 166–179 (2017)