The Impact of Preprocessing Pipeline Choice in Univariate and Multivariate Analyses of PET Data

Martin Nørgaard[1,2], Douglas N. Greve[5], Claus Svarer[1], Stephen C. Strother[4], Gitte M. Knudsen[1,2], Melanie Ganz[1,3]
[1]*Neurobiology Research Unit, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark*
[2]*Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
[3]*Department of Computer Science, University of Copenhagen, Copenhagen, Denmark*
[4]*Rotman Research Institute at Baycrest, University of Toronto, Toronto, Canada*
[5]*Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA*

***Abstract*—It has long been recognized that the data preprocessing chain is a critical part of a neuroimaging experiment. In this work we evaluate the impact of preprocessing choices in univariate and multivariate analyses of Positron Emission Tomography (PET) data. Thirty healthy participants were scanned twice in a High-Resolution Research Tomography PET scanner with the serotonin transporter (5-HTT) radioligand [^{11}C]DASB. Binding potentials (BP_{ND}) from 14 brain regions are quantified with 384 different preprocessing choices. A univariate paired t-test is applied to each region and for each preprocessing choice, and corrected for multiple comparisons using FDR within each pipeline. Additionally, a multivariate Linear Discriminant Analysis (LDA) model is used to discriminate test and retest BP_{ND}, and the model performance is evaluated using a repeated cross-validation framework with permutations. The univariate analysis revealed several significant differences in 5-HTT BP_{ND} across brain regions, depending on the preprocessing choice. The classification accuracy of the multivariate LDA model varied from 37% to 70% depending on the choice of preprocessing, and could reasonably be modeled with a normal distribution centered at 51% accuracy. In spite of correcting for multiple comparisons, the univariate model with varying preprocessing choices is more likely to generate false-positive results compared to a simple multivariate analysis model evaluated with cross-validation and permutations.**

I. Introduction

Positron Emission Tomography (PET) is an invaluable tool used in many aspects of state-of-the-art neuroscience to capture the spatiotemporal distribution of neurotransmitters and receptors in the brain. However, due to limitations in data acquisition, the generative signals making up these PET images are significantly affected by complex spatiotemporal noise patterns, consequently resulting in a suboptimal signal-to-noise ratio (SNR). These limitations have led to the development of a large array of data preprocessing strategies designed to remove artefacts and noise from the images. It has long been recognized that preprocessing is a critical part of the PET analysis framework, with new PET radioligands often being required to have been carefully validated in a test-retest setting with different kinetic models and at different scan lengths (Parsey et al. 2000, Ginovart et al. 2001). Nonetheless, several subsequent studies deviate substantially from these analyses and guidelines presented in published validation studies, implicitly assuming that the chosen set of preprocessing steps are insensitive to the outcome measure and produce near-optimal results (Nørgaard et al. 2018). Despite the importance and usefulness of validating kinetic models and scan length, the impact of several other important factors such as preprocessing strategies for delineating volumes of interest (VOI), whether to apply motion correction (MC), how to accurately perform co-registration, and whether to use partial volume correction (PVC), remains unclear. In this study, we will extend the question of the influence of preprocessing choices to also include the subsequent statistical analysis using either univariate of multivariate analysis approaches. This is important because the statistical analysis largely depends on the quality of the data going into the analysis, and may therefore produce biased and non-reproducible results if the uncertainty of the data is not taken into account.

II. Methods

A. PET and MRI Data Collection

All participants were scanned using a Siemens ECAT High-Resolution Research Tomography (HRRT) scanner operating in 3D list-mode and with the highly selective radioligand [^{11}C]DASB. The imaging protocol consisted of a single-bed, 90 minutes transmission acquisition post injection of 587 $\pm$ 30 (mean $\pm$ SD) MBq, range 375-612 MBq, bolus into an elbow vein. PET data was reconstructed into 36 frames (6x10, 3x20, 6x30, 5x60, 5x120, 8x300, 3x600 seconds) using a 3D-OSEM-PSF algorithm with TXTV based attenuation correction (image matrix, 256 x 256 x 207; voxel size, 1.22 x 1.22 x 1.22 mm) (Sureau et al. 2008, Keller et al. 2013). PET data was obtained from 30 healthy women (mean age: 25 $\pm$ 5.9 years, range: 18 - 37) from a previous randomized, placebo-controlled and double-blind intervention study investigating the role of 5-HTT changes in depressive responses to sex-steroid hormone manipulation (Frokjaer et al. 2015). The women served as a control group receiving placebo only, i.e., the data represent test-retest without any expected changes in [^{11}C]DASB binding. All participants were PET scanned two times with a median interval of 34 days (range: 27 - 122 days). An anatomical 3D T1-weighted MP-RAGE sequence with matrix size = 256 x 256 x 192; voxel size = 1 x 1 x 1 mm; TR/TE/TI = 1550/3.04/800 ms; flip angle = 9°

978-1-5386-4291-7/18/$31.00 ©2018 IEEE

was acquired for all participants using a Siemens Magnetom Trio 3T MR scanner or a Siemens 3T Verio MR scanner. Additional information can be found in Frokjaer et al. 2015. The study was registered and approved by the local ethics committee (protocol-ID: H-2-2010-108). All participants gave written informed consent.

B. Preprocessing

We evaluated the effects of applying a sequence of five preprocessing steps to the PET data, followed by either a univariate or multivariate analysis model. The final outcome measure for each pipeline is the non-displaceable binding potential (BP_{ND}) in 14 representative brain regions: *amygdala, thalamus, putamen, caudate, anterior cingulate cortex (ACC), hippocampus, orbital frontal cortex, superior frontal cortex, occipital cortex, superior temporal gyrus, insula, medial-inferior temporal gyrus, parietal cortex, and entorhinal cortex*. Each preprocessing step consisted of 2-4 choices, and all the choices have previously been used in the PET literature. The steps are listed below in the order in which they were applied, combinatorially summing to a total of 384 preprocessing pipelines.

1. Delineation of Volumes of Interest (VOI): All MRI scans were processed using FreeSurfer (FS) (http://surfer.nmr.mgh.harvard.edu, version 5.3). Subsequently to running the FS pipeline, manual edits can be applied to correct for errors. If a T2-weighted MRI is available, semi user-independent edits can be made to the FS output by re-running the FS pipeline with the T2-weighted MRI. We examined all three choices, and now refer to these as FS-RAW (standard output), FS-MAN (output with manual edits) and FS-T2P (output with the T2 stream).

2. Motion correction (MC): PET MC was executed using AIR (v. 5.2.5). Prior to alignment, each frame was smoothed using a 10 mm Gaussian 3D kernel and thresholded at the 20-percentile level. Alignment parameters were estimated for PET frame 10-36 using AIR, geometrically transformed using a scaled least squares cost-function, and resliced into a 4D motion corrected data set (Frokjaer et al. 2015). The data was analyzed either with or without MC.

3. Co-registration: All single-subject PET time activity curves (TACs) were initially either summed or averaged over all time frames to estimate a time-weighted (twa) or averaged (avg) 3D image for co-registration. Two different co-registration techniques were subsequently applied to either the twa or the avg image, namely Normalized Mutual Information (NMI, Studholme et al. 1999) or Boundary-Based Registration (BBR, Greve et al. 2009). This results in four choices for co-registration.

4. Partial Volume Correction (PVC): The data were analyzed either without or with three different partial volume correction (PVC) approaches. The VOI-based PVC technique, Geometric Transfer Matrix (GTM), by Rousset et al. 1998 was applied, establishing a forward linear model relating [^{11}C]DASB intensities to the VOI means, as described in Greve et al. 2016. Because the PSF for a HRRT scanner varies from 1-4 mm depending on the distance from the center of the field-of-view (Olesen et al. 2009), we ran the analyses with the PSF settings; 0 mm, 2 mm, and 4 mm.

5. Kinetic Modeling (KinMod): We applied four kinetic modeling approaches, all based on reference tissue modeling (RTM). These include the Multilinear Reference Tissue Model (MRTM) and the Multilinear Reference Tissue Model 2 (MRTM2) by Ichise et al. 2003. The non-invasive Logan reference tissue model was applied as described in Logan et al. 1996, and the Simplified Reference Tissue Model, SRTM, was applied as described by Lammertsma and Hume, 1996.

C. Univariate Analysis

The difference in estimated BP_{ND}'s between test and retest sessions as a function of pipeline *J* and region *K*, was evaluated using paired t-tests. All data was tested for normality using a Kolmogorov-Smirnov (KS) test. Within each pipeline, *J*, the 14 regions were corrected for multiple comparisons using False Discovery Rate (FDR, Benjamini & Hochberg) at $q = 0.05$. A P-value less than 0.05 is considered a significant result and represents a false positive.

D. Multivariate Analysis

In this study, we used a multivariate Linear Discriminant Analysis (LDA) model for predictive classification of test (class 1) and retest (class 2) BP_{ND}. For this two-class dataset, $\mathbf{X} \in \mathbb{R}^{14}$, LDA estimates an optimal discriminant that maximizes the ratio of between-class covariance to within-class covariance. We can write the conditional posterior probability of $\mathbf{X}$ originating from class C_k as the following:

$$p(\mathbf{X}|C_k;\theta) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}||\mathbf{L_{train}}^T(\mathbf{X} - \bar{\mathbf{X}}^k_{\mathbf{train}})||^2\} \quad (1)$$

where $\bar{\mathbf{X}}^k_{\mathbf{train}}$ is the training data mean from class C_k, and $\mathbf{L_{train}}$ is a linear transformation matrix normalized so that training variance is unity. From (1), we can estimate the posterior probability of correct class assignment $p(C_k|\mathbf{X};\theta)$. The model was trained by subsampling 80% of the data (balanced data-set of 24 test and 24 re-retest scans) in a 5-fold cross-validation framework. The model was then evaluated using a validation set, $\mathbf{X}$, consisting of the remaining 20% (6 subjects with test and re-test scans). The validation data was independent of the training data and completely held out of the training procedure. The subsampling procedure was repeated so that each label was assigned to the validation data exactly once. The entire cross-validation framework was repeated 10 times to obtain an unbiased mean classification accuracy (Varoquaux et al. 2017). The significance of each model was estimated by randomly permuting the class labels 1000 times and re-running the above 10 times repeated 5-fold cross-validation procedure to generate an empirical null-distribution. This provides an empirical P-value for each model and pipeline.

III. Results

The classification accuracy is estimated as the number correctly classified labels divided by the total number of labels.

A. Univariate Analysis

The paired t-test was applied to the entire dataset (i.e. test and retest BP_{ND}) and for the 384 pipelines. The false positive rates (FPR) are summarized in Figure 1 and 2 for the uncorrected and corrected for multiple comparisons using FDR, respectively, with higher FPR being worse.

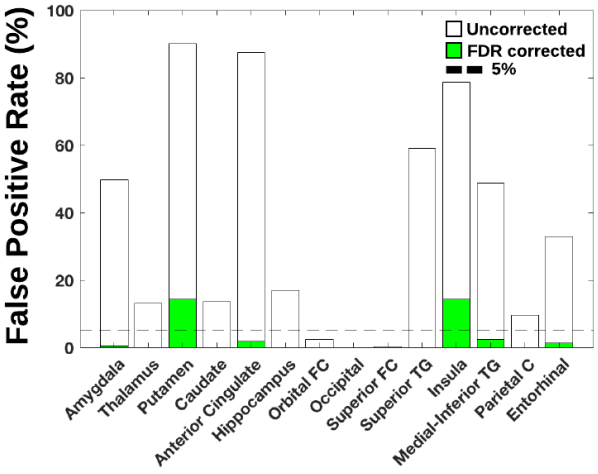

Fig. 1. Number of significant results (paired t-test, $P < 0.05$) in 384 pipelines divided by 384, expressed as a percentage for 14 brain regions. Blank is not corrected for multiple comparisons, whereas green is corrected using FDR.

All significant results reported passed the KS test. The uncorrected analysis shows a large percentage of significant results (1929 out of 5376 statistical tests) for both subcortical and cortical regions (Figure 1). When correcting for multiple comparisons using FDR, the number of significant results is dramatically reduced to 133 significant results (Figure 2). However, for several brain regions, significant results can still be obtained and are influenced by different choices in the preprocessing pipeline (Figure 2). In general, the choices of preprocessing being mostly responsible for the significant results (i.e. false positive results) are MC, and the kinetic models MRTM and SRTM.

B. Multivariate Analysis

The results of the multivariate analysis are presented in Figure 3A and 3B for the preprocessing-dependent and permuted classification accuracies, respectively. Depending on the choice of preprocessing, the classification accuracy varied from 37% to 70% across all repetitions, with a mean accuracy and standard deviation of 51% and 4%, respectively. The pipeline that produced the highest classification accuracy (maxPipeline) was: VOI=FS-T2P, MC=no, Co-reg=NMI_{AVG}, PVC=no, KinMod=MRTM. The mean accuracy for this pipeline was 63.3% ($P = 0.12$) relative to the randomly permuted distribution. One of the 10 repetitions of the 5-fold cross-validation for maxPipeline produced a classification

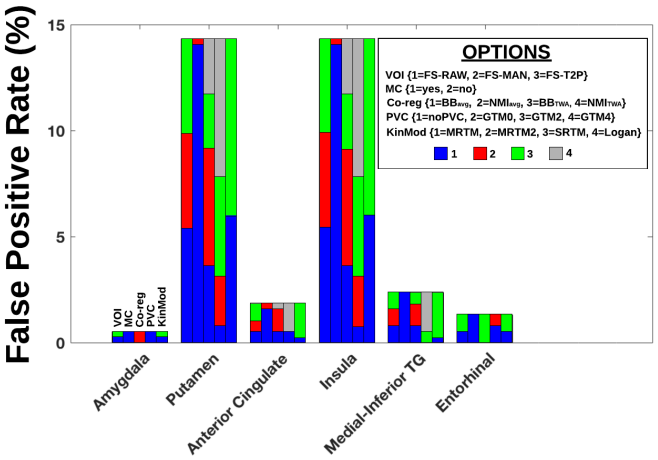

Fig. 2. Number of significant results (paired t-test, $P < 0.05$) in 384 pipelines divided by 384, expressed as a percentage for 14 brain regions (corrected for multiple comparisons at FDR=0.05 within each pipeline). The five vertical bars within each region represent the distribution of choices, and have the order: 1. VOI (1=FS-RAW, 2=FS-MAN, 3=FS-T2P), 2. MC (1=yes, 2=no), 3. Co-reg (1=BB_{avg}, 2=NMI_{avg}, 3=BB_{twa}, 4=NMI_{twa}), 4. PVC (1=noPVC, 2=GTM0, 3=GTM2, 4=GTM4), 5. KinMod (1=MRTM, 2=MRTM2, 3=SRTM, 4=Logan).

accuracy of 70%, and thereby significantly different from its permuted null-distribution at $P = 0.01$ (Figure 3B).

IV. Discussion

Here, we present a comprehensive framework for testing the impact of a wide range of preprocessing pipeline choices in combination with univariate and multivariate analysis models. The presented results question the validity of preprocessing pipeline choices being independent of the neuroimaging outcome in [^{11}C]DASB measurements using PET. For univariate models without correction for multiple comparisons, the percentage of significant results was largely inflated (36% significant results across all pipelines and regions) given the experimental design being a test-retest study with no expected changes between scans. When correcting for multiple comparisons using FDR, several significant results were still present. In a post-hoc analysis, we also corrected the results using Bonferroni correction within each pipeline, producing a total of 23 significant results in putamen ($N = 1$) and insula ($N = 22$) across all pipelines. This corresponds to 0.4% significant results with Bonferroni compared to 2.5% with FDR, across 5376 statistical tests.

Regarding the performance of the multivariate models, the distinction between test and retest BP_{ND} as a function of preprocessing pipeline choice was not evident. We illustrate that the spread of classification accuracies as a function of preprocessing pipeline (Figure 3A) can reasonably be modeled as a Gaussian signal distribution with mean 51% and standard deviation 4%. Notably, the significant classification finding for a single cross-validation run depicted in Figure 3B suggests that, depending on the preprocessing choice and without performing repeated cross-validation, significant results (i.e. false positives) are obtained using a multivariate model and

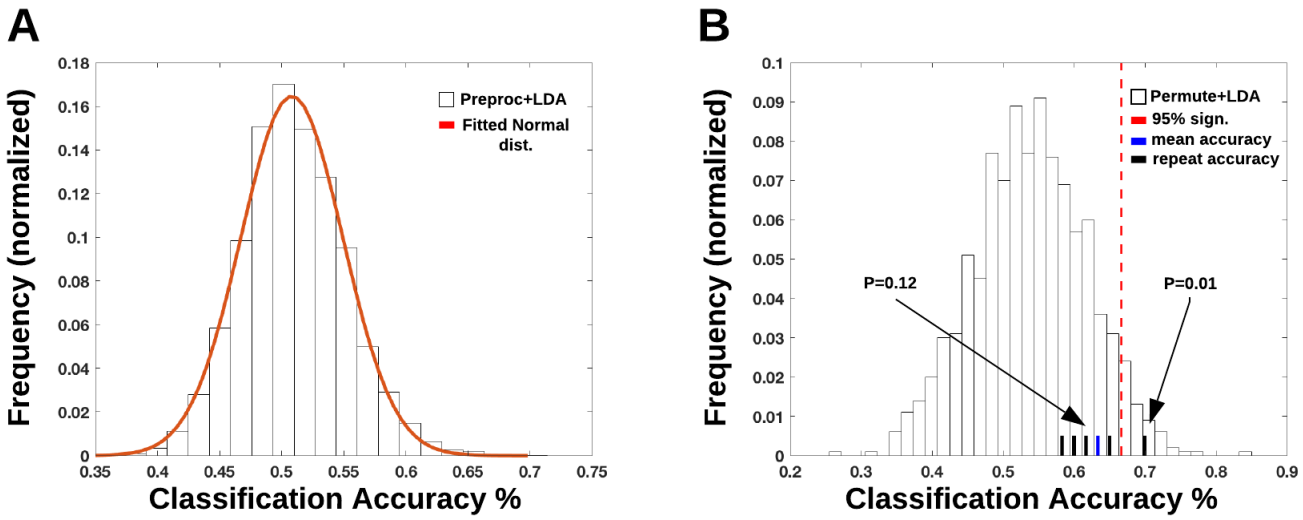

Fig. 3. **(A)** Normalized distribution of classification accuracies (%) for 10 times repeated 5-fold cross-validation and for 384 different preprocessing choices **(B)** Normalized distribution of 1000 permuted classification accuracies (%) for the pipeline maximizing the classification accuracy in (A). The black bars are the classification accuracy for 10 individual repetitions for the pipeline and the blue bar is the mean classification accuracy over the 10 repetitions. One of the repetitions by chance produces a classification accuracy higher than the 95% significance level (red vertical dotted line).

with permutations. This is simply due to the variance in the cross-validation results. This behaviour was also described in detail by Varoquaux et al. 2017, advocating to perform repeated cross-validation and to use the mean as an unbiased estimator of classification performance.

A. Future Work

The performance of univariate and multivariate analysis models as a function of preprocessing pipeline should optimally be evaluated for all radiotracers. While there can be several reasons for why we observe a difference between test and retest, ranging from biological biases, data acquisition biases and preprocessing biases, it becomes non-trivial how we can subsequently separate these components (Kim et al. 2006). These potential biases can be added as variables in future models to explain variation, however, this quickly becomes an ill-posed problem given the high dimensionality of the data and low sample sizes. A limitation of our test-retest study is that there could be a possible order and/or placebo effect present. This has not been reported previously and warrants further investigation.

REFERENCES

[1] Nørgaard M, Ganz M, Svarer C, et al. Cerebral Serotonin Transporter Measurements with [^{11}C]DASB: A Review on Acquisition and Preprocessing across 21 PET Centres. Journal of Cerebral Blood Flow and Metabolism, 2018. Accepted.

[2] Parsey RV, Slifstein M, Hwang DR, et al. Validation and reproducibility of measurement of 5-HT1A receptor parameters with [carbonyl-^{11}C]WAY-100635 in humans: comparison of arterial and reference tissue input functions. J Cereb Blood Flow Metab 2000, 20(7):1111-1133.

[3] Ginovart, N., Wilson, a. a., Meyer, J. H., Hussey, D., and Houle, S. (2001). Positron emission tomography quantification of [^{11}C]DASB binding to the human serotonin transporter: modeling strategies. Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism, 21(11):1342-1353.

[4] Frokjaer, V. G., Pinborg, A., Holst, K. K., et al. Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: A positron emission tomography study. Biological Psychiatry 2015, 78(8):534-543.

[5] Studholme C, Hill DLG, and Hawkes DJ. An overlap invariant entropy measure of 3d medical image alignment. Pattern Recogn. 32, pp. 71-86, 1999.

[6] Greve D, Fischl B (2009). Accurate and robust brain image alignment using boundary-based registration. Neuroimage, 48, 63-72.

[7] Rousset, O.G.,Ma, Y., Evans, A.C., 1998. Correction for partial volume effects in PET: principle and validation. J. Nucl. Med. 39, 904-911.

[8] Olesen, O. V., Sibomana, M., Keller, S. H., et al. Spatial resolution of the HRRT PET scanner using 3D-OSEM PSF reconstruction. IEEE Nuclear Science Symposium Conference Record, 2009, 3789-3790.

[9] Greve, D. N., Salat, D. H., Bowen, S. L., et al. Different partial volume correction methods lead to different conclusions: An 18F-FDG-PET study of aging. NeuroImage 2016, 132:334-343.

[10] Ichise, M., Liow, J.-S., Lu, J.-Q., et al. Linearized reference tissue parametric imaging methods: application to [^{11}C]DASB positron emission tomography studies of the serotonin transporter in human brain. Journal of Cerebral Blood Flow and Metabolism 2003: Official Journal of the International Society of Cerebral Blood Flow and Metabolism, 23(9), 1096-1112.

[11] Logan J, Fowler JS, Volkow ND, et al. Distribution volume ratios without blood sampling from graphical analysis of PET data. J Cereb Blood Flow Metab 1996, 16(5):834-840.

[12] Lammertsma, A.A., Hume, S.P., 1996. Simplified reference tissue model for PET receptor studies. Neuroimage 4, 153-158.

[13] Varoquaux G, Raamana PR, Engemann D, et al. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage. Volume 145, Part B, 15 January 2017, Pages 166-179.

[14] Sureau FC, Reader AJ, Comtat C, et al. Impact of image-space resolution modeling for studies with the high-resolution research tomograph. J Nucl Med, 2008; 49(6): 1000-8.

[15] Keller SH, Svarer C, Sibomana M. Attenuation correction for the HRRT PET-scanner using transmission scatter correction and total variation regularization. IEEE Trans Med Imaging, 2013; 32(9): 1611-21.

[16] Kim, J. S., Ichise, M., Sangare, J., et al. PET Imaging of Serotonin Transporters with [^{11}C]DASB: Test-Retest Reproducibility Using a Multilinear Reference Tissue Parametric Imaging Method. J. Nucl. Med. 2006, 47(2), 208-214.