UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

REGION H

Rigshospitalet

Optimizing Preprocessing Pipelines in PET/MR Neuroimaging

PhD Thesis

Martin Nørgaard

Copenhagen, January 29th, 2019

Main Supervisor
Prof. Dr. Gitte Moos Knudsen, MD, DMSc
Neurobiology Research Unit,
Rigshospitalet, University of Copenhagen

Co-Supervisors
Senior Scientist, Claus Svarer, PhD
Neurobiology Research Unit,
Rigshospitalet, University of Copenhagen
Prof. Stephen C. Strother, PhD
Department of Medical Biophysics,
University of Toronto
Asst. Prof. Melanie Ganz, PhD
Department of Computer Science,
University of Copenhagen

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen on January 29th, 2019

Neurobiology Research Unit
Rigshospitalet
Copenhagen University Hospital

Neurobiology Research Unit
Rigshospitalet, University of Copenhagen
Blegdamsvej 9, DK-2100 Copenhagen, Denmark
Phone +45 35454215, Fax +45 35454015
rigshospitalet@rh.regionh.dk
`www.rigshospitalet.dk`

MGH/HST Martinos Center for Biomedical Imaging
Laboratory for Computational Neuroimaging
149 Thirteenth Street, Charlestown, MA 02129
`https://www.martinos.org/lab/lcn`

Rotman Research Institute Baycrest
Strother Lab
3560 Bathurst Street, Toronto, Ontario, Canada M6A 2E1
Phone 416-785-2500
`http://www.baycrest.org/`

Submitted in fullfillment of the requirements for the degree of

Doctor of Philosophy

at
University of Copenhagen

Main Supervisor
Prof. Gitte M. Knudsen, Neurobiology Research Unit, Rigshospitalet, University of CPH

Co-supervisors
Senior Scientist, Claus Svarer, PhD, Neurobiology Research Unit, Rigshospitalet
Prof. Stephen C. Strother, PhD, Dept. of Medical Biophysics, University of Toronto
Asst. Prof. Melanie Ganz, PhD, Dept. of Computer Science, University of Copenhagen

Assessment Committee
Prof. Liselotte Højgaard, Dept. of Clinical Physiology, Rigshospitalet, University of CPH
Prof. Ronald Boellaard, PhD, VU University Medical Center, Amsterdam, NL
Prof. R. Todd Ogden, PhD, Columbia University, New York, US

Abstract

Positron Emission Tomography (PET) is a state-of-the-art imaging technique for measuring the spatial distribution of neurotransmitters and receptors in the living human brain. However, the PET signal is influenced by complex spatio-temporal noise patterns arising from sources of radioactive decay, head motion and scanner-specific limitations. A large set of preprocessing algorithms have been developed to remove various sources of noise, but there is currently a limited consensus in the literature on the most optimal preprocessing strategy. Furthermore, it is not well understood how the choice of preprocessing strategy may affect the variability of the data and ultimately the conclusions of a study. This thesis develops a framework for the evaluation of preprocessing performance in PET using the radioligand [^{11}C]DASB, targeting the serotonin transporter, as exemplary case. In the five included research papers, I evaluate current preprocessing strategies in the literature, how they affect measures of test-retest bias, variability and false-positive rates, and how they may lead to different conclusions in a double blind, randomized, placebo-controlled study. Finally, I provide a statistical framework for adequately controlling the false-positive rate when dealing with large sets of preprocessing options.

In this work, I show that (1) variations in choice of preprocessing strategy are an overlooked aspect in modern PET neuroscience, (2) measures of bias, within- and between-subject variability are significantly affected by preprocessing strategy, and significant differences between test and retest were obtainable despite correcting for multiple comparisons and (3) different preprocessing strategies lead to different neurobiological conclusions. My findings suggest that the preprocessing stage contributes with considerable variance into the data, with the preprocessing steps motion correction, partial volume correction and kinetic modeling contributing the most. I show that knowledge about the variability of preprocessing is critical to limiting false-positive rates. This underlines the importance of selecting preprocessing strategy with great caution. Finally, I present my view on future directions and best practices for handling preprocessing variability across PET centres.

Resume in Danish

Positron Emissions Tomografi (PET) er en medicinsk billeddannende teknik til at måle biokemiske og farmakologiske processer i den levende menneskehjerne. Der er imidlertid stigende bekymring over, hvor vanskeligt det har været at replikere denne type forskning, og meget tyder på, at støjkilder fra optagelsen af PET data, samt valget af hvorledes data forbehandles (præ-processeringen) har afgørende betydning for det endelige resultat. En lang række præ-processerings strategier er gennem årene blevet udviklet til at fjerne støjkilder, men der er uenighed omkring valget af den mest optimale strategi. Derudover fremgår det ikke klart, hvorledes valget af præ-processering påvirker variabiliteten i data, og dermed hvilke konklusioner, der kan drages. I denne afhandling udvikler jeg en strategi, hvormed man baserer sine valg af præprocesserings-trin på kvantitative mål, ved anvendelse af data optaget med radioliganden [^{11}C]DASB. I de fem inkluderede artikler viser jeg først, hvor meget valget af præ-processering varierer i literaturen. Dernæst viser jeg, hvorledes valget af præ-processering påvirker variabiliteten og falsk-positiv raten i et test-retest datasæt, samt hvordan det påvirker konklusionerne i et randomiseret, placebo-kontrolleret studie. Afslutningsvist, udvikler jeg et statistisk redskab til at kontrollere for falsk-positiv raten, når der eksisterer mange muligheder for valg af præ-processerings strategier. Mine resultater viser, (1) der er stor variation i literaturen omkring valg af præ-processering (2) statistiske mål som bias, variabilitet (i samme person og i mellem personer), samt falsk-positiv raten påvirkes betydeligt af præ-processering, og (3) forskellige valg af præ-processering resulterer i forskellige konklusioner. Mine resultater demonstrerer, at præ-processering bidrager med betydelig variabilitet i data, hvor præ-processerings valg: bevægelses-korrektion, partial volume korrektion og kinetisk modellering, er de komponenter, der bidrager mest. Jeg demonstrerer også, at viden om variabiliteten af præ-processering er kritisk for a mindske falsk-positiv raten. Dette understreger vigtigheden af, at valg af præ-processering skal baseres på grundig analyse og tilpasses det biologiske spørgsmål. Afslutningsvist bidrager jeg med mit syn på fremtidig forskning, samt bedste fremgangsmåder til at håndtere præ-processering på tværs af PET centre.

Preface

This thesis was prepared during a two and a half year period (from August 2016 to February 2019) at the Neurobiology Research Unit, Copenhagen University Hospital, Rigshospitalet, including a 5 month external research stay at the MGH/Harvard-MIT Martinos Center for Biomedical Imaging, Boston, USA. The thesis was submitted in conformity with the requirements for the degree of Doctor of Philosophy at University of Copenhagen.

The thesis deals with evaluation and optimization of preprocessing strategies for PET/MRI neuroimaging.

The thesis was handed in on January 29th 2019

Supervisors
Prof. Gitte Moos Knudsen, DMSc, University of Copenhagen, Rigshospitalet
Senior Scientist, Claus Svarer, PhD, Rigshospitalet
Prof. Stephen C. Strother, PhD, Baycrest Hospital, University of Toronto
Asst. Prof. Melanie Ganz, Dept. of CS, University of Copenhagen

Assessment Committee
Prof. Liselotte Højgaard, DMSc, University of Copenhagen, Rigshospitalet
Prof. Ronald Boellaard, VU University Medical Center, Amsterdam, NL
Prof. R. Todd Ogden, Columbia University, NY, USA

Copenhagen, January 29th, 2019

Martin Nørgaard

Papers included in the thesis

Publications

[A] **Nørgaard M**, Ganz M, Svarer C, Feng L, Ichise M, Lanzenberger R, Lubberink M, Parsey RV, Politis M, Rabiner EA, Slifstein M, Sossi V, Suhara T, Talbot PS, Turkheimer F, Strother SC, Knudsen GM. 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*, 2019 Feb;39(2):210-222.
DOI: 10.1177/0271678X18770107.

[B] **Nørgaard M**, Ganz M, Svarer C, Frokjaer VG, Greve DN, Strother SC, Knudsen GM. Optimization of Preprocessing Strategies in Positron Emission Tomography (PET) Neuroimaging: A [^{11}C]DASB Study. In revision, NeuroImage, Nov 2018.

[C] **Nørgaard M**, Greve DN, Svarer C, Strother SC, Knudsen GM, Ganz M. The Impact of Preprocessing Pipeline Choice in Univariate and Multivariate Analyses of PET Data. *Pattern Recognition in Neuroimaging (PRNI)*, IEEE Explore, 2018, pp. 1-4. DOI: 10.1109/PRNI.2018.8423962

[D] **Nørgaard M**, Ganz M, Svarer C, Greve DN, Frokjaer VG, Strother SC, Knudsen GM. The Impact of Different Preprocessing Strategies in PET Neuroimaging: A [^{11}C]DASB-PET Case. Submitted to *Journal of Cerebral Blood Flow and Metabolism*, Jan 2019.

[E] **Nørgaard M**, Ozenne B, Svarer C, Frokjaer VG, Ganz M. Preprocessing, Prediction and Significance: Framework and Application to Brain Imaging. Submitted to Medical Image Computing and Computer Assisted Intervention (MICCAI), Jan 2019.

Other Relevant Publications

1 **Nørgaard M**, Ganz M, Svarer C, Beliveau V, Fisher PM, Mc Mahon B, Greve DN, Strother SC, Knudsen GM. Estimation of Regional Seasonal Variations in SERT-levels using the FreeSurfer PET pipeline: a reproducibility study. *Proc. of the MICCAI workshop on Computational Methods for Molecular Imaging*, 2015. In press.

2 Rasmussen JH, **Nørgaard M**, Hansen AE, Vogelius IR, Aznar MC, Johannesen HH, Costa J, Kjær A, Specht L, Fischer BM. Feasibility of multiparametric imaging with PET/MR in head and neck squamos cell carcinoma. Journal of Nuclear Medicine, 2017: 58(1): pp. 69-74. DOI: 10.2967/jnumed.116.180091

3 Deen M, Hansen HD, Hougaard A, da Cunha-Bang S, **Nørgaard M**, Svarer C, Keller SH, Thomsen C, Ashina M, Knudsen GM. Low 5-HT1B receptor binding in the migraine brain: A PET study. Cephalalgia. 2018 Mar;38(3):519-527. DOI: 10.1177/0333102417698708

4 **Nørgaard M**, Ganz M, Svarer C, Fisher PM, Churchill NW, Beliveau V, Grady C, Strother SC, Knudsen GM. Brain Networks Implicated in Seasonal Affective Disorder: A Neuroimaging PET Study of the Serotonin Transporter. Frontiers in Neuroscience | Brain Imaging Methods, November 2017. DOI: 10.3389/fnins.2017.00614

5 Deen M, Hansen HD, Hougaard A, **Nørgaard M**, Eiberg H, Lehel S, Ashina M, Knudsen GM. High brain serotonin levels in migraine between attacks: A 5-HT4 receptor binding PET study. Neuroimage: Clinical 18 (2018); 97-102.

6 Mc Mahon B, **Nørgaard M**, Svarer C, Andersen SB, Madsen MK, Baare W, Madsen J, Frokjaer VG, Knudsen GM. Seasonality-resilient individuals

downregulate their cerebral 5-HT transporter binding in winter - A longitudinal combined 11C-DASB and 11C-SB207145 PET study. European Neuropsychopharmacology, 2018, Oct;28(10):1151-1160.

Acknowledgements

First of all, I would like to give a sincere thank you to my supervisor and mentor, Professor Gitte Moos Knudsen. Thank you, Gitte, for taking me under your wings and for giving me the opportunity to be a part of a fantastic research group. Your innate dedication to science, your close to infinite knowledge of neuroscience, and your always positive attitude and leadership are a true inspiration. Finally, thank you for all our discussions on all aspects of science, and for always believing in me and pushing me to become a better researcher. I would also like to thank my other supervisors, Asst. Prof. Melanie Ganz, Senior scientist Claus Svarer, and Prof. Stephen C. Strother. Thank you, Melanie, for your ability to always motivate and inspire researchers around you, for your engagement in this project, and for always wanting the best for me. Thank you, Claus, for letting me benefit from your solid experience and knowledge, for all your stories, and for always taking the time to discuss topics with me ranging from preprocessing and statistics, to international history and weather phenomena's. Thank you, Stephen, for sharing your in exhaustive knowledge on preprocessing and statistics, and for always taking good care of me. Especially, thank you for showing me to balance between science and enjoying life (travelling, food, wine), and for showing me that this balance is important for doing exemplary science. I would also like to thank Asst. Prof. Douglas Greve and the Martinos Center for Biomedical Imaging at MGH/Harvard/MIT in Boston for kindly hosting me during a five-month stay. Thank you, Doug, for all our discussions on preprocessing, coding and mathematical derivations, and for teaching me to always be critical.

I would also like to thank everyone at the Neurobiology Research Unit (NRU) at Rigshospitalet. Special thanks to Hanne Demant for being absolutely awesome, and for all our conversations, runs, and good times over the years. Thank you to my office mates, Sebastian Holst, Ling Feng, Vincent Beliveau and Giske Opheim, for all your positive energy, all the laughs, and for being absolutely amazing at every Christmas party. Thank you to the remaining part of the data analysis team, Patrick Fisher, Brice Ozenne, and Martin Schain for all your constructive feedback and for a lot of good discussions over the years. Thank you to my workout buddy, colleague and friend Martin Korsbak, for many good times and jokes, and for doing really cool research. Thank you to Marie Deen for all our discussions on migraine and preprocessing, and for being excellent company in Boston. A sincere thank you to both Dorthe Givard and Peter S. Jensen for always helping me out on practical matters and for extracting data from the CIMBI database.

Thank you to Vibe G. Frøkjær for letting me use her high quality data, and for being subject to a full investigation of her data with respect to preprocessing and statistical analysis. Also thank you to Marie Deen and Brenda Mc Mahon for including me in their research on migraine and Seasonal Affective Disorder, respectively, and for always creating a great atmosphere, whether it being at a party or at NRU. Thank you to all the remaining people at NRU for creating a pleasant and fun research environment, for all the laughs, and last but not least, for all the cake!

Finally, I would like to express my sincerest thanks to my family (Keld, Lene, Anne, Jeremy, Nadine, Anna Marie) and to the love of my life, Kathrine Kielsgaard, for your love and support and for understanding my high commitment to this project.

Martin Nørgaard, Copenhagen, January 29th 2019

Nomenclatures

Abbreviations

AAL	Automated Anatomical Atlas
AD	Alzheimers Disease
ACC	Anterior Cingulate Cortex
Acc	Accuracy
AIR	Automated Image Registration
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
AVG	Average
BBR	Boundary Based Registration
BDNF	Brain-derived neurotrophic factor
BGO	Bismuth germanate detector
Bq	Becquerel
BP_{ND}	Non-displaceable binding potential
BSV	Between-Subject Variability
CNR	Contrast-to-Noise-Ratio
CI	Confidence Interval
CSF	Cerebrospinal Fluid
CV	Coefficient of Variation
DASB	3-amino-4-(2-dimethylaminomethylphenylsulfanyl)-benzonitrile

FBP	Filtered Back Projection
FC	Frontal Cortex
FDG	^{18}F-FluoroDeoxyGlucose [MBq]
FDR	False Discovery Rate
FIX	Optimal Fixed Pipeline
fMRI	Functional Magnetic Resonance Imaging
FOV	Field-Of-View [cm^2]
FPR	False-Positive Rate
FS	FreeSurfer
FWHM	Full-Width-Half-Maximum
GE	General Electric
GLM	General Linear Model
GnRH	Gonadotropin Releasing Hormone
gSNR	Global Signal-To-Noise Ratio
GSO	Gadolinium oxyorthosilicate detector
GTM	Geometric Transfer Matrix
HRRT	High Resolution Research Tomograph
ICC	Intraclass Correlation Coefficient
KS	Kolmogorov-Smirnov
LDA	Linear Discriminant Analysis
LOR	Line-Of-Response
LSO	Cerium-doped lutetium oxyorthosilicate detector
LYSO	Lutetium-yttrium oxyorthosilicate detector
MDD	Major Depressive Disorder
MNI	Montreal Neurological Institute
MP-RAGE	Magnetization-Prepared Rapid Gradient-Echo
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MRTM	Multilinear Reference Tissue Model
MRTM2	Multilinear Reference Tissue Model 2
NaI(TI)	Thallium-doped sodium iodide detector
NEC	Noise Effective Counts
nMC	without Motion Correction

noPVC	without Partial Volume Correction
OSEM	Ordered Subset Expectation Maximization
PET	Positron Emission Tomography
PSF	Point Spread Function
PVC	Partial Volume Correction
PVE	Partial Volume Effect
RTM	Reference Tissue Model
SAD	Seasonal Affective Disorder
SNR	Signal-to-Noise Ratio
SPM	Statistical Parametric Map
SRTM	Simplified Reference Tissue Model
SSRI	Selective Serotonin Reuptake Inhibitor
SUV	Standardized Uptake Value $[\frac{Bq}{kg}]$
TAC	Time Activity Curve
TF	Tissue Fraction
TWA	Time Weighted Average
VOI	Volume Of Interest $[cm^3]$
WM	White Matter
WSV	Within-Subject Variability
5-HT	Endogenous Serotonin
5-HTT	Serotonin Transporter
5-HTTLPR	Serotonin-transporter-linked polymorphic region

Symbols

$\mathbf{X}$	Design matrix
$\bar{\mathbf{X}}_{\mathbf{train}}$	Training data mean from class C_k
$\mathbf{L}_{\mathbf{train}}$	Linear transformation matrix normalized to unit variance C_k
C_k	Class assignment k
y	Column vector of $[^{11}C]$DASB uptake values
y_n	Class labels $\in \{-1, 1\}$
g	Column vector of class labels
β	Row vector with true underlying uptake values
$\hat{\beta}$	Estimate of the VOI means

k_2	Tracer clearance rate in target tissue [min^{-1}]
$k_2^{'}$	Tracer clearance rate in reference tissue [min^{-1}]
$C(T)$	Tissue tracer concentration in target region at time T [kBq/mL]
$C^{'}(T)$	Tissue tracer concentration in reference region at time T [kBq/mL]
V	Total distribution volume in target region [mL/mL]
$V^{'}$	Total distribution volume in reference region [mL/mL]
t	Time [s]
b	Intercept term
$\otimes$	Convolution
R_1	Tracer delivery of relative influx of tracer to target
BP_{ND}	Non-Displaceable Binding Potential
DVR	Distribution Volume Ratio
S	Number of sub-samples
$\tilde{n}$	Number of subjects drawn from statistical subsampling
$\hat{n}$	Estimate of number of needed subjects
J	Number of preprocessing pipelines
K	Number of regions
N	Number of samples
M	Number of repeats cross-validation
d	Difference in binding between test and retest
μ	Group average of binding
σ	Group standard deviation of binding
E	Effective change in binding (sample size)
R	Pearson's correlation
Z	Number of permutations
Θ	Heaviside step function
$\hat{P}_{max}$	Cumulative distribution of max accuracies
$\mathbf{\Pi}$	Set of all permutations
π	Permutation sample from a uniform distribution
p	Number of features
δt	Timing window in PET imaging [ms]

Contents

Abstract **i**

Resume in Danish **iii**

Preface **v**

Papers included in the thesis **vii**

Acknowledgements **xi**

Nomenclatures **xiii**

1 Introduction **1**

2 Motivation and Background **5**
- 2.1 Positron Emission Tomography 5
  - 2.1.1 Principles of PET 5
  - 2.1.2 Measuring Radioligand Binding with Dynamic PET . . . 8
- 2.2 The Data-Analysis Chain in Dynamic PET 9
  - 2.2.1 Subject Selection 10
  - 2.2.2 Data Acquisition 10
  - 2.2.3 Preprocessing 11
  - 2.2.4 Statistical Analysis 15
- 2.3 Optimization of Preprocessing Strategies 17
  - 2.3.1 Performance Metrics 18
- 2.4 Thesis Objectives 19

3 Methods - The Data-Analysis Chain **21**
- 3.1 Subjects 21

3.2 Data Acquisition 22
3.3 Preprocessing 22
3.4 Statistical Analysis 27
3.4.1 Performance Metrics of Reproducibility 27
3.4.2 Statistical Analysis Models 29

4 Study 1: Preprocessing Strategies in the PET Literature 31
4.1 Introduction 31
4.2 Methods 32
4.2.1 The Data-Analysis Chain 32
4.3 Results 33
4.3.1 Subject Selection 33
4.3.2 Data Acquisition 34
4.3.3 Preprocessing 35
4.3.4 Statistical Analysis 35
4.4 Discussion 36

5 Study 2: Evaluation and Optimization of Preprocessing 39
5.1 Introduction 39
5.2 Methods 40
5.2.1 The Data-Analysis Chain 40
5.2.2 Preprocessing Optimization Across Subjects and Regions 41
5.2.3 Region-Specific Preprocessing Optimization 41
5.3 Results 42
5.3.1 Test-retest Bias 42
5.3.2 Within- and Between-Subject Variability 43
5.3.3 Preprocessing Optimization Across Subjects and Regions 44
5.3.4 Region-Specific Preprocessing Optimization 45
5.3.5 False-Positive Analysis 46
5.4 Discussion 48

6 Study 3: Different Preprocessing Choices Lead to Different Conclusions 51
6.1 Introduction 51
6.2 Methods 52
6.2.1 The Data-Analysis Chain 52
6.3 Results 53
6.3.1 Depressive Symptoms and Preprocessing 53
6.3.2 Neuroticism and Preprocessing 54
6.4 Discussion 55

7 Study 4: Predictive Framework to Correct for Multiple Preprocessing Options 59
7.1 Introduction 59

7.2 Methods 61
    7.2.1 Defining a Subset of Preprocessing Strategies 62
    7.2.2 Model Selection and Cross-validation 62
    7.2.3 Permutation Test for a Single Strategy 62
    7.2.4 Permutation Test for Multiple Strategies 63
7.3 Experiments 64
7.4 Discussion 65

8 Conclusions and Future Work 67
8.1 Summary 67
8.2 Future Work 68
    8.2.1 Voxelwise and Surface-Based Preprocessing 68
    8.2.2 False-Positive Rates in PET 69
    8.2.3 Data Sharing 70
8.3 Conclusions 71

Bibliography 71

A Paper [A] 91

B Paper [B] 105

C Paper [C] 147

D Paper [D] 153

E Paper [E] 173

F Declarations of Co-Authorship 183

Chapter 1

Introduction

Positron Emission Tomography (PET) is a state-of-the-art neuroimaging technique for imaging receptor systems (e.g. dopamine or serotonin) *in vivo.*
PET provides 4D imaging of the entire brain with relatively good spatial/temporal resolution (millimeters/seconds), and with high sensitivity/specificity for molecular targets (pico molar concentrations). It is a unique tool in neuroscience for studying drug effects in the living human brain, but also expands to a broader range of clinical applications such as the detection of cancerous tissue [Boellaard et al., 2015, Fischer et al., 2009], evaluation of myocardial perfusion and metabolism [Kero et al., 2017, Danad et al., 2014], and for quantifying the progression of Alzheimer's Disease (AD) [Zwan et al., 2017, Cohen and Klunk, 2014].

Serotonin is a neurotransmitter critical to homoeostasis, and its regulation and timing are important determinants of health [Azmitia, 1999]. Insufficient regulation of serotonin has been associated with a range of brain disorders including depression, anxiety disorders, sleep disturbance, attention deficit disorder, schizophrenia and AD, all together constituting the largest socioeconomic burden in Western societies [Wittchen et al., 2011]. Although our understanding of the serotonin system has advanced in recent years, several findings have been contradictory, characterized by an inability to produce, and reproduce, reliable biomarkers of disease risk and treatment responsiveness. This may, in part, stem from an incomplete understanding of the sources of variation in the acquired data. However, while most published receptor studies using PET mainly have focused on extracting neuroscientifically relevant results, only a limited number of studies have investigated the extent to which these findings may be influenced by different sets of preprocessing steps ('preprocessing pipeline/stage') applied when analyzing the data. A preprocessing pipeline in neuroimaging commonly refers to a set of steps used to denoise and remove artifacts in the data for subsequent statistical analysis (e.g. motion correction and outlier detection), thereby improving the overall quality of the data. PET centres or even individual scientists often design their own unique preprocessing strategy, and as a result, there is currently no consensus in the PET community on the most optimal preprocessing strategy. This is further complicated by the fact that preprocessing is not carried out in isolation, but rather depends on several other stages in a PET workflow (the "Data-Analysis Chain", Figure 1.1) with various parameter choices, each of which may interact with preprocessing to influence the signal and noise. These stages include subject heterogeneity (Step 1), PET data acquisition (Step 2), and choice of statistical analysis model (Step 4).

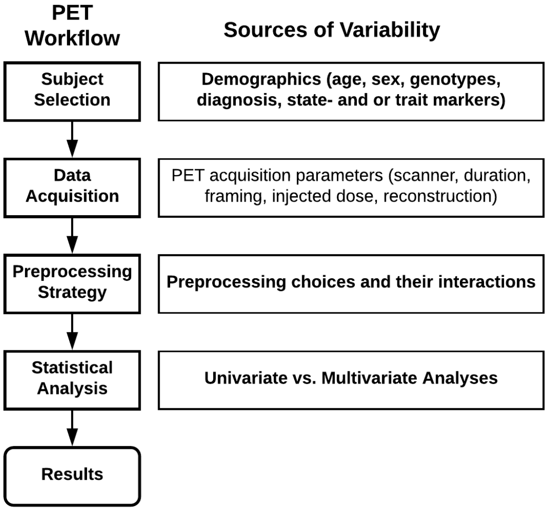
PET Workflow
Sources of Variability
Subject Selection
Demographics (age, sex, genotypes, diagnosis, state- and or trait markers)
Data Acquisition
PET acquisition parameters (scanner, duration, framing, injected dose, reconstruction)
Preprocessing Strategy
Preprocessing choices and their interactions
Statistical Analysis
Univariate vs. Multivariate Analyses
Results

Figure 1.1: Workflow in a common PET experiment. The workflow consists of 4 major stages: 1) subject selection, 2) data acquisition, 3) preprocessing, 4) statistical analysis. Choices at each stage may significantly affect the signal and noise, but may also interact to influence the results.

Differences in receptor-occupancy as measured by PET are characterized by relatively weak and non-stationary signal changes, typically ranging between 5-20% following pharmarcological intervention (e.g. [Jørgensen et al., 2018]), and with complex sources of structured noise. The principal noise components in PET are typically subject-dependent, including head motion effects and physiological processes, such as respiration and cardiac pulsation [Reyes et al., 2007, Lamare et al., 2007]. Studies suggest, that motion artefacts are present in 10-20% of high-resolution PET data [Ooi et al., 2009]. Furthermore, accompanying noise confounds are additionally amplified during long acquisition scans [van der Kouwe et al., 2006, Kober et al., 2012], especially in cases where patients suffer from medical conditions preventing them from staying still in the scanner [Aksoy et al., 2011, Andrews-Shigaki et al., 2011, Forman et al., 2011]. The signal changes caused by such confounds are highly variable between subjects, and the integration of complex temporal and spatial signals making up these data, challenges a reliable interpretation in studies with low sample sizes [Button et al., 2013]. To reduce subject-specific artefacts, a broad range of preprocessing algorithms have been developed, ranging from de-noising (e.g. spatial smoothing) to artefact-specific correction (e.g. partial volume correction or motion correction). It is commonly assumed in PET that there exists a single preprocessing strategy that can be adapted to all subjects to produce optimal results. However, it has not been well explored how individual subjects or groups of subjects are heterogeneous in their optimal preprocessing strategy. In addition, there is evidence supporting the notion that preprocessing demands (e.g. motion correction) vary as a function of other stages in a PET experiment, such as data acquisition [Boellaard et al., 2001] and statistical analysis [Fisher et al., 2017], although these issues need further validation.

Taken together, there is a need for a quantitative framework for evaluating and comparing the performance of preprocessing strategies in PET, and a need to test potential preprocessing interactions with subject variability and choice of statistical analysis. In the following section (Chapter 2) I will provide the foundation of this thesis, namely the motivation and background. This includes the principles behind PET, including the extension to dynamic PET and measurement of radioligand binding. Then, I will review the stages in the Data-Analysis Chain of a typical dynamic PET experiment, ranging from subject selection to the final results, and review how they relate to preprocessing. This is preceded by a discussion on strategies for preprocessing optimization. Finally, I will explicitly state the research objectives of this thesis in detail.

Thesis overview: The thesis includes a background and motivation part (Chapter 2). The background chapter introduces the reader to the main topics and limitations of a PET experiment, and strategies for optimization and validation of preprocessing pipelines. Chapter 2 is rounded off with the research objectives of this thesis. Chapter 3 contains the methods of the PET Data-Analysis Chain used in this thesis, including details to evaluate and optimize preprocessing pipelines. Chapters 4-7 cover the main results of this thesis (studies 1-4), presenting the five scientific articles from the Appendices A-E. A thesis conclusion is found in Chapter 8, including a perspective on future work.

Chapter 2

Motivation and Background

2.1 Positron Emission Tomography

2.1.1 Principles of PET

Positron Emission Tomography (PET) is a quantitative nuclear imaging technique, in which the emission of positrons from the nucleus of a radioactive atom is used to construct molecular images. When positrons interact with electrons, they are annihilated, causing two 511 KeV photons to emit linearly in almost opposite direction (Figure 2.1). The line in which the photons are emitted is known as a Line Of Response (LOR). The process of forming a LOR is fundamental in nuclear medicine, where a radioactive isotope is either injected or inhaled into the body. The isotope will distribute throughout the body by blood circulation, and accumulate in specific tissue cells depending on the biochemical structure of the radiotracer. Here the radioactive isotope will emit positrons as it decays. The two emitted photons can be detected in coincidence using gamma ray detectors and the signal can subsequently be converted into an electrical signal, amplified, and reconstructed into a 3D image containing the spatial location of the decay. The theory behind PET, including some of its limitations, can be summarized as the following:

After the positron leaves the nucleus it will have an initial kinetic energy. However, due to elastic and inelastic interactions with surrounding matter, it will eventually lose its kinetic energy making the distance travelled from the nucleus finite. The finite distance travelled contributes to uncertainties from where the radioactive decaying nucleus originated. This is rather essential as the main purpose in PET is to estimate the location of the decaying nucleus and not the location of the annihilation. In addition, not all photon pairs are emitted strictly at an angle of 180°. In water only 35% of the annihilations have zero momentum and emit photons exactly at an angle of 180°. This combined with the positron traveling distance before annihilation are some of the limitations affecting the resolution of a PET scanner. For an isotope such as ^{11}C the positron mean travel range in water is approximately 1 mm [Bailey et al., 2005, p. 22].

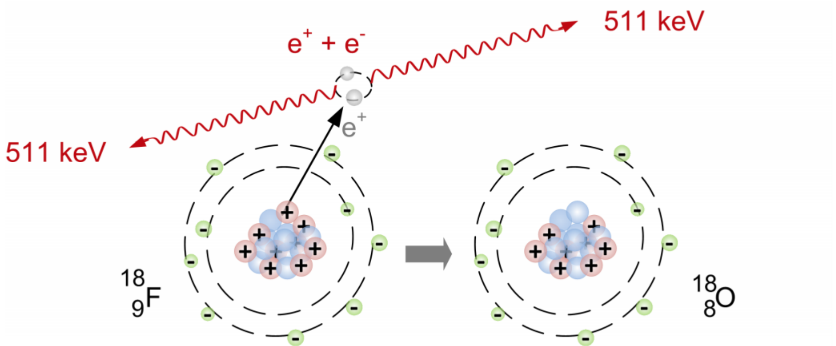

Figure 2.1: Annihilation as a result of a positron being ejected from the nucleus of a $^{18}_{9}$F atom. The annihilation occurs due to positron-electron merging, thereby creating two photons to be sent off in almost linear opposite direction.

Attenuation of Radiation and Interaction with Matter

High-energy photons can interact with matter in three different ways; the photoelectric effect, Compton scattering and pair production [Bailey et al., 2005]. The extent to which the photons interact with surrounding matter is predominantly determined by the energy of the photon, and the corresponding matters ability to absorb energy. In the *photoelectric effect* the photon will collide with a bound electron of an atom and transfer all of its energy to the electron. This will subsequently result in the emission of an electron from the atom. In *Compton scattering* the photon interacts with a loosely bound orbital electron of an atom and will transfer only a part of its kinetic energy. The loosely bound electron will subsequently be ejected from the atom, and the photon will be scattered in a new direction with an angle related to its loss of energy. *Compton scattering* occurs frequently within the human body at an energy interval of approximately 100 keV to 2 MeV. *Pair production* is the third option for photons to interact with matter. Here a high-energy photon with kinetic energy higher than 1.022 MeV collides with a surrounding nucleus.

Several crystals can be used in PET imaging (e.g. NaI(T), BGO, LSO, LYSO, GSO) to detect the photons and they are all characterized by having different physical properties. Mainly four properties for crystals are essential for its proper application in PET; stopping power for 511 keV photons, signal decay time, light output, and intrinsic energy resolution [Bailey et al., 2005].

Coincidence Detection

Subsequently to annihilation (coincidence event) the two emitted photons are detected by two scintillators, but in order to measure the spatial point of annihilation a timing window needs to be introduced in this context. A timing window can be defined as being a short time interval, δt, for the detection of photons from a coincident event within the field-of-view (FOV). Mainly, coincidence events are divided into three categories: random, true and scatter. *Random coincidences* occur when two photons from independent events are detected at two opposing detectors and within the coincidence timing window (10^{-8}sec). Random coincidences are therefore not representing a single coincident event, and they mainly appear as a result of a too large timing window. Random coincidences and scatter are mainly the two most detrimental effects in PET imaging that need to be corrected for in order to get *true coincidences* only (Figure 2.2). To further reduce the contribution of randoms and scatter, it is also common in PET to select an energy window, constraining the photons to lie within the range of typically 400 KeV to 600 KeV.

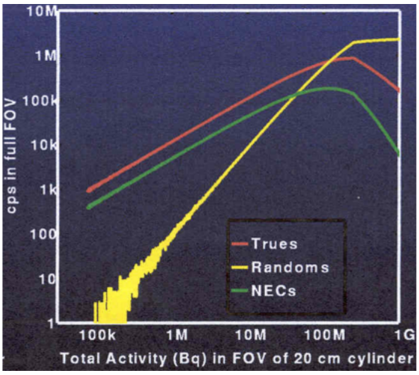

Figure 2.2: Counts per second (cps) and the association with the total activity (Bq). Red is the true counts, yellow is the random counts, and green is the Noise Effective Counts (NEC) ([Holm et al., 1995]).

Spatial Resolution

The contribution from all the above limitations (traveling distance before annihilation, interaction with matter, and detector principles) make the resulting PET image susceptible to Partial Volume Effects (PVE). The PVE causes the radiotracer signal from a point object to have a "spread out", appearing larger than it actually is. This blur effect is caused by the spatial resolution of the scanner (Point Spread Function, PSF), a quantitative measure of how well a PET scanner can differentiate between two objects in close vicinity. The reconstructed PET image can be approximated by assuming that the true underlying image has been volume-smoothed with a Gaussian kernel at a known resolution (PSF). This measure is called the full width at half maximum (FWHM) and is equal to the spatial resolution of the scanner. As a consequence of this, a small FWHM equals a small spatial resolution. In this sense, smaller is therefore better.

2.1.2 Measuring Radioligand Binding with Dynamic PET

Dynamic PET studies that can measure radioligand uptake over time *in vivo* are increasingly receiving attention in the field of neuroimaging due to their high specificity at the receptor-level. Dynamic PET studies measure the distribution of a radioligand over sequential time intervals, whereas studies using static PET measure the distribution over a single time interval, hence providing no temporal information. The dynamic measurements can be reconstructed into a sequence of 3D images (frames) that contain the concentration of radioactivity (Bq/mL) as a function of time (time-activity curve, TAC) from each voxel (volume element) or region (contiguous set of voxels).

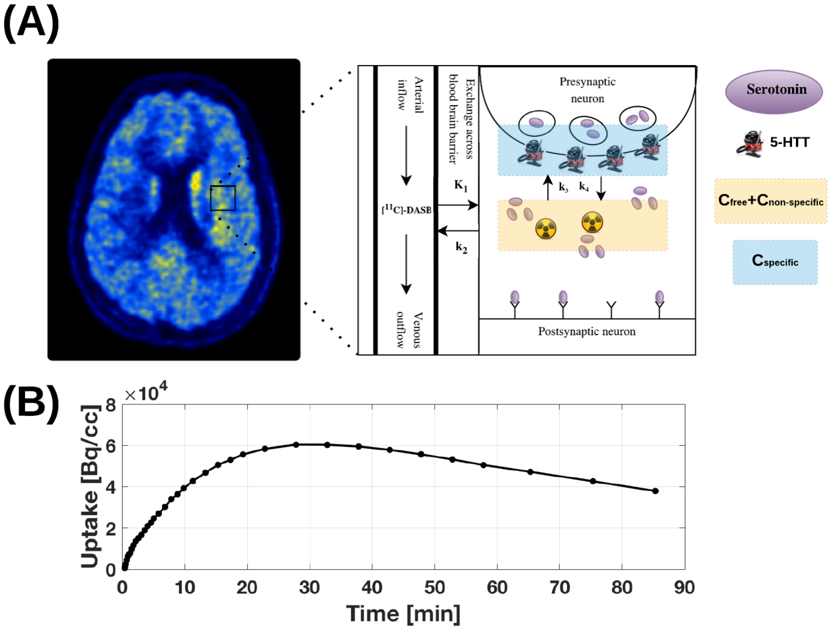

Figure 2.3: **(A)** Overview of the modeling assumptions regarding tracer delivery, uptake, binding, and clearance of a radioligand in a single voxel **(B)** Time Activity Curve (TAC) for the voxel in (A) depicting the total distribution of radioligand over time.

The time-varying distribution of radioligand can be used to mathematically model the physiological parameters of interest such as perfusion and receptor densities. The model seeks to explain the kinetic behaviour of the radioligand by introducing a number of possible compartments. For example, a radioligand targeting the serotonin transporter (5-HTT) may be specifically bound in the synapse or it may distribute freely without binding to 5-HTT (Figure 2.3A). The modeling will be covered in detail in the preprocessing section 2.2.3.

2.2 The Data-Analysis Chain in Dynamic PET

PET experiments typically consist of complex workflows, with multiple stages ranging from (1) subject selection, (2) experimental design, (3) data acquisition, (4) preprocessing, (5) statistical analysis to the neurobiological interpretation (Figure 2.4). However, choices made at any stage in a PET workflow may significantly affect the signal and noise in the data. Furthermore, the stages are not independent from each other and may interact to influence the results. The optimization of a PET workflow is often performed with the aim of optimizing only a single stage and/or step, leaving other variables fixed. However, in order to optimize a PET workflow, it is important to have deep knowledge of each step and empirically examine how the steps may interact to influence the results. The details of each stage are outlined below with a special focus on the use of the radioligand [^{11}C]DASB, targeting the 5-HTT. Furthermore, at each stage, an extra emphasis is put on the interaction with preprocessing and how it may influence the results of the analyses. The focus of this thesis is the interaction between subject selection (stage 1), preprocessing strategy (stage 4) and statistical analysis (stage 5), although the interactions between preprocessing and the other stages are also reviewed.

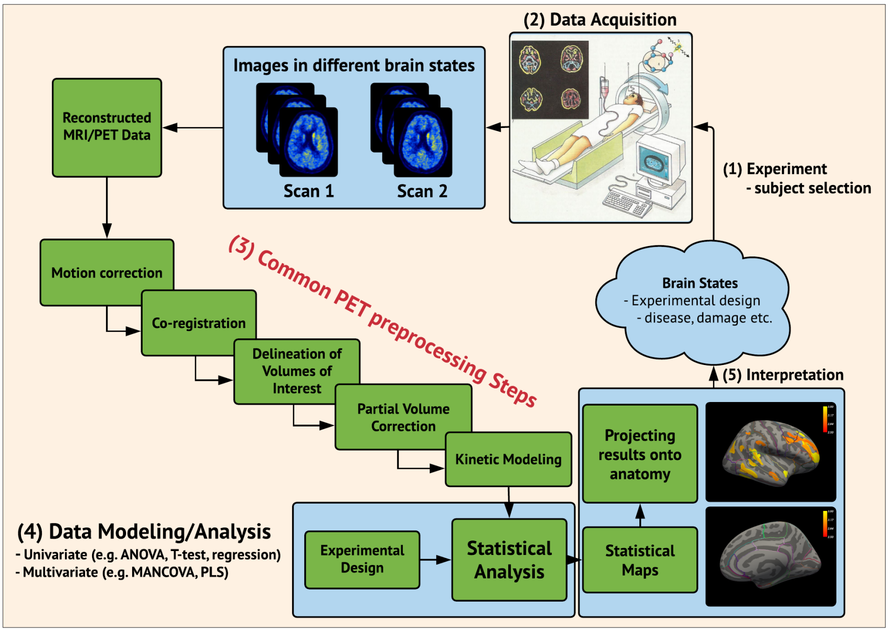
(2) Data Acquisition
Images in different brain states
Scan 1
Scan 2
Reconstructed MRI/PET Data
(1) Experiment
- subject selection
Motion correction
Co-registration
Delineation of Volumes of Interest
Partial Volume Correction
Kinetic Modeling
(3) Common PET preprocessing Steps
Brain States
- Experimental design
- disease, damage etc.
(5) Interpretation
Projecting results onto anatomy
Statistical Maps
(4) Data Modeling/Analysis
- Univariate (e.g. ANOVA, T-test, regression)
- Multivariate (e.g. MANCOVA, PLS)
Experimental Design
Statistical Analysis

Figure 2.4: Flowchart depicting a common pipeline for neuroimaging studies (multimodal PET and MRI) and its multiple stages ranging from (1) experimental design / subject selection, (2) data acquisition, (3) preprocessing, (4) data modeling/analysis, and (5) interpretation.

2.2.1 Subject Selection

Extensive research in humans supports the notion that 5-HTT densities, as can be measured with [^{11}C]DASB-PET, are subject-dependent and may vary as a function of age, sex, genotype (5-HTTLPR or BDNF val66met) and stress-levels [Fisher et al., 2017, Cannon et al., 2006, Kalbitzer et al., 2010]. While these latter components may all contribute to variation directly at the receptor level, other types of variation may appear in terms of subject-specific head motion, respiration, cardiac pulsations and diurnal variation. A number of studies have shown that head motion significantly affects the signal-to-noise (SNR) ratio and renders the PET data disturbed or even useless [Anton-Rodriguez et al., 2010, Green et al., 1994]. Variations in respiration and cardiac pulsations will also affect the blood flow, and thereby the delivery of radioligand to the brain tissue. Furthermore, changes in blood flow may influence the neuronal response and vascular coupling, but this effect has been shown for a few receptor-systems to have limited impact (e.g. [Sander et al., 2019]). Diurnal variation has been reported to result in lower 5-HTT levels across the day [Matheson et al., 2015], although only males were included in this latter cross-sectional study. Other factors that have been reported to affect the serotonin system include seasonal changes [Mc Mahon et al., 2016, Nørgaard et al., 2017], variations in menstrual cycle [Jovanovic et al., 2009] and personality traits such as neuroticism [Tuominen et al., 2017] and anxiety [Cannon et al., 2006].

The preprocessing strategy that optimizes signal detection may vary substantially between subjects even for subjects characterized as homogeneous [Zanderigo et al., 2017]. As it is expected that patterns of subject-specific noise will vary across subjects, it is also expected that the degree to which the noise can be removed using a fixed preprocessing strategy will vary across subjects. For example, partial volume correction (PVC) is recommended in studies where brain atrophy interacts with an effect of interest (e.g. age or diagnosis), and failure to properly account for partial volume effects in these cases can falsely inflate or degrade the effect of interest [Greve et al., 2016]. Limited understanding exists how subject-specific sources of variation and their potential interactions may influence the subsequent acquisition of the PET signal as well as other parts of the data-analysis chain.

2.2.2 Data Acquisition

Data acquisition in dynamic PET studies typically involve a combination of implicit and explicit parameter choices that can be tuned and optimized to control the signal and noise. The explicit parameters include total scanning time, injected dose, specific activity, attenuation correction and the use of head masks to reduce subject-specific motion [McMahon et al., 2018, Ogden et al., 2007].

The implicit parameters (pertaining to the scanner hardware/software) include time window length, energy window, spatial resolution, sensitivity, framing and the reconstruction [Morimoto et al., 2006, Belanger et al., 2004, Boellaard et al., 2001]. The implicit and explicit parameters may interact with each other to affect both signal and noise, and there are important trade-offs to be made that have been demonstrated to affect subsequent steps in the data-analysis chain [Anton-Rodriguez et al., 2010, Green et al., 1994, Jin et al., 2014]. However, as this thesis mainly focuses on preprocessing interactions with statistical analyses, a comprehensive examination of data acquisition is considered beyond the scope and needs to be examined in future work. Nevertheless, as several important components of the data acquisition may influence the preprocessing and statistical analyses, I here provide a brief overview of some of the interactions and limitations.

The SNR in PET increases with injected dose, but there is global optimum to optimize for (Figure 2.2) where the fraction of random counts will catch up with the fraction of true counts to reduce the Noise Effective Counts (NEC) [Holm et al., 1995]. High-resolution scanners have higher sensitivity compared to clinical scanners but they are limited by smaller detector elements, resulting in more noise due to scatter and head motion [Wienhard et al., 2002, van Velden et al., 2009]. The size and dimensions of the detector elements have important trade-offs with increasing size of the detector resulting in increased spatial resolution [van Velden et al., 2009]. In addition, some scanners favour axial resolution over transaxial resolution (e.g. GE Advance PET scanner, [Khohlmyer and Stearns, 2002]). This latter parameter choice produces non-isotropic spatial resolution, resulting in different spill-over effects (partial volume effects, PVEs) of radiotracer in different directions. A number of components in the PET acquisition may contribute to PVEs including detector properties, traveling distance to annihilation, head motion and the reconstruction algorithm. These choices will interact with subsequent preprocessing steps such as PVC and spatial smoothing [Greve et al., 2014].

2.2.3 Preprocessing

Preprocessing in dynamic PET commonly refers to a set of algorithms used to denoise and remove artifacts in the data for subsequent statistical analysis (e.g. motion correction and PVC), thereby improving the overall quality of the data. In dynamic PET, this typically includes 5 main categories: (1) motion correction (2) co-registration (3) delineation of volumes-of-interest (4) PVC and (5) kinetic modeling for quantification of radioligand binding. Although it has been suggested that there exists nearly as many unique analysis pipelines in the literature as there are studies [Carp, 2012b], this thesis specifically focuses on the subset of preprocessing strategies that are most common for dynamic PET, with the exception of the analysis of arterial blood data. All the included preprocessing steps have various tuning parameters that can be optimized to control signal and noise. The details of these parameter choices are discussed below.

Motion Correction

Motion correction (MC) is typically performed as the first preprocessing step in dynamic brain PET studies, with the goal of removing head motion artefacts induced during the data acquisition (Figure 2.5). The most common application of MC is frame-by-frame correction, where alignment parameters for each frame are (1) estimated to a reference frame by minimizing a cost function, (2) transformed using the estimated alignment parameters and (3) resliced into a 4D motion corrected data set. It has been shown by numerous paper that head motion in PET brain imaging renders PET data disturbed or even useless [Olesen et al., 2013, Anton-Rodriguez et al., 2010, Green et al., 1994]. [Freire and Mangin, 2001], and [Orchard and Atkins, 2003], demonstrated that least-squares cost functions may be susceptible to activation biases and outliers, which for PET means that the MC algorithm may attempt to incorrectly account for motion if the image has low SNR, or if the tracer distribution in the target volume substantially changes over time compared to the reference volume. Conversely, in the absence of motion, MC will lead to some degree of smoothing due to an interpolation in the reslicing. Frame-by-frame motion correction without re-doing the image reconstruction may also result in errors in attenuation correction in the PET reconstruction, which is often neglected [van den Heuvel et al., 2003].

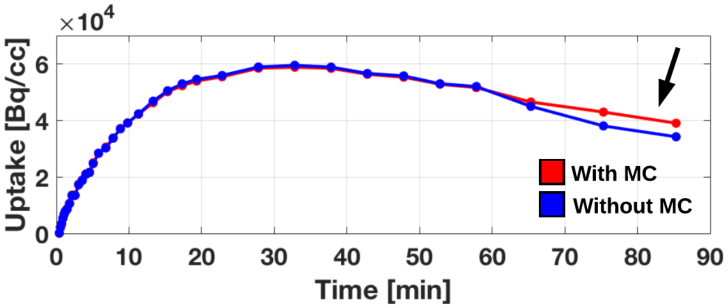

Figure 2.5: Example showing the Time Activity Curves of [^{11}C]DASB-PET uptake in the thalamus either with motion correction (red) or without (blue).

Co-registration

Co-registration is typically performed in PET studies as a rigid-body volume transformation of a reference image to a target image. The target image is often a structural Magnetic Resonance Image (MRI) containing anatomical information or a group-atlas with predefined volumes-of-interest. A precise co-registration is important, as voxels that move across tissue boundaries are susceptible to extreme signal changes [Schwarz et al., 2017]. The cost-functions used for co-registration are similar to as for MC. However, as the reference image (PET) and target image (MRI) have different spatial resolution and spatial scales (voxel size), resampling with spatial interpolation is always carried out. The spatial interpolation is commonly used to boost SNR, at the expense of image resolution [Strother et al., 2004].

Delineation of Volumes-of-Interest

Many PET studies are driven by hypotheses related to specific anatomical brain structures, often referred to as volumes-of-interest (VOIs). For PET this generally requires co-registration to a structural MRI with anatomically labeled regions (atlas), as the signal in PET does not reflect anatomical information. However, publicly available atlases have different VOIs of the same biological region varying in both size, location and delineation technique (Figure 2.6). Some studies provide evidence that there is good agreement between certain atlases [Schain et al., 2014], whereas other studies suggest substantial variations between atlases [Nørgaard et al., 2015].

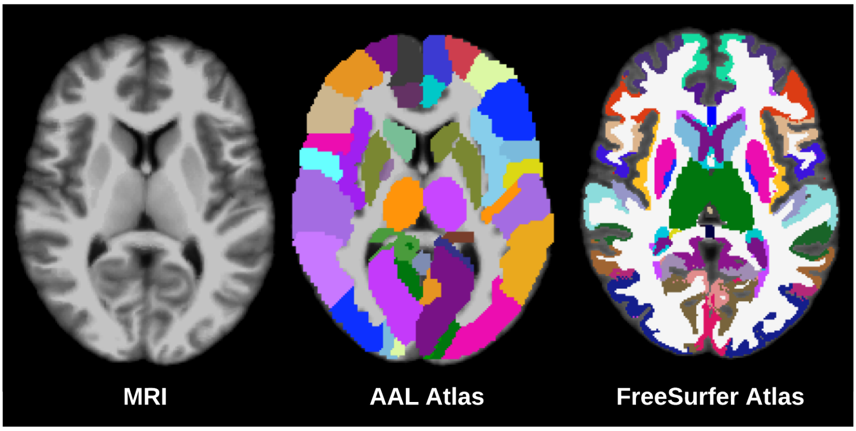

Figure 2.6: Structural MRI (left), overlayed with delineated regions from the Automated Anatomical Labeling (AAL) atlas (center), and overlayed with delineated regions from the FreeSurfer atlas (right).

Manual labeling, as opposed to automatic, may impose an interrater bias in the data, unless well-defined operational criteria and blindness to diagnosis are enforced. Another potential issue with both manual delineations and atlases is the assumption of homogeneously distributed tracer within the region. If this assumption is violated it will misrepresent the true underlying radioligand concentration within that region.

Partial Volume Correction

Partial Volume Correction (PVC) is typically performed by (1) estimating the spill-in and spill-out of signal between tissue types with different neuronal properties (i.e. gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF)) and then (2) removing the contribution of spill-in while simultaneously compensating for spill-out signal. The degree of spill-over effects is largely determined by the PSF of the scanner, and an increased spatial resolution will result in an increased degree of partial volume effects. Furthermore, as the PSF not only varies between PET scanners but also on the distance from the centre of the field of view [Olesen et al., 2009], it is important to make proper assumptions about the PSF when performing PVC so it matches the location and resolution of the VOI. As different tissue types have different partial volume effects, a homogeneous and accurate segmentation of each tissue type is required [Greve et al., 2016]. However, the utility of PVC is disputed. It has been shown to cause noise amplification [Rousset et al., 2007] and reduce measurement bias [Schwarz et al., 2018]. However, it has also been found to potentially induce a systematic bias, reflecting subject-dependent differences in anatomy and not true differences in radioligand uptake [Greve et al., 2016]. PVC is typically recommended in studies where brain anatomy interacts with an effect of interest (e.g. age or diagnosis), and failure to properly account for PVEs in these instances can falsely impact the results [Müller-Gärtner et al., 1992, Meltzer et al., 1999, Greve et al., 2016]. [Greve et al., 2016] suggested the Geometric Transfer Matrix (GTM) to be the preferred method for VOI analysis compared to no PVC, but this method largely depends on a homogeneous distribution of tracer inside the VOI. The method therefore remains to be fully validated as the preferred method. Nevertheless, many PVC methods have been criticized for being subject to arbitrary selection of parameter choices, consequently resulting in limited consensus in the literature on the importance and/or use of PVC [Greve et al., 2016].

Kinetic Modeling

For kinetic modeling using reference tissue models (RTM), the final output is the non-displaceable binding potential ($\mathrm{BP}_{\mathrm{ND}}$) for each given region [Innis et al., 2007]. RTMs rely on the assumption and identification of a reference region with non-specific binding characteristics. In the [^{11}C]DASB-PET literature the cerebellum has commonly been used as a reference region because of its absence of 5-HTT. However, the use of cerebellum as a reference region is questionable. Some researchers argue for the use of cerebellum [Ginovart et al., 2001], whereas others argue against [Miller et al., 2016], as [^{11}C]DASB BP_{ND} has been shown to be displaced following treatment with selective serotonin reuptake inhibitors (SSRIs) [Parsey et al., 2006b]. RTMs mainly dissociate from each other in the model-parameter estimation varying from linear (MRMT2; [Ichise et al., 2003]) to non-linear techniques (SRTM; [Lammertsma and Hume, 1996]). The methods also vary in terms of assumptions and how the noise is controlled (e.g. MRTM vs. MRTM2, [Ichise et al., 2003]), and how many parameters are necessary to fit the data (two parameters (MRTM2) vs. three parameters (SRTM and MRTM)). However, there is a bias-variance trade-off to consider, as a reduction in number of parameters to fit the data will reduce the variance of the model, at the expense of a bias [Ichise et al., 2003]. It is however, largely unknown how this bias-variance trade-off impacts the quantification in individual subjects and regions, and hence may influence the results in a group analysis.

2.2.4 Statistical Analysis

Once preprocessed, the goal of a dynamic PET study is often to establish group differences in region-specific BP_{ND} between a group with a medical condition (e.g. depression) and a healthy control group. While this study design is cross-sectional, it may also be expanded to a longitudinal design, where participants are scanned more than once (e.g., [Mc Mahon et al., 2016]). The study may also include an intervention [Frokjaer et al., 2015], scores of depression or any other external variables that may be used as confounds or correlates with the BP_{ND}. The end result is a statistical measure reflecting the association between the BP_{ND} (dependent variable) and the external variables (independent variables). The statistical analysis is often carried out using two general approaches, namely univariate or multivariate analysis techniques.

Univariate analysis models are limited to a single dependent variable, and individual brain regions of specific binding are therefore assumed to be independent, consisting of randomly-sampled mixtures of signal and noise. Univariate analysis models are often linear in nature, e.g. ANOVA (Analysis of Variance), ANCOVA (Analysis of Covariance), linear or rank correlation analyses (Pearson, Kendall or Spearmann) and t-tests. Together these linear models constitute special cases of the General Linear Model (GLM, [Friston et al., 1995]) generalizing multiple linear regression to the case of p dependent variables [Chen et al., 2014, Monti, 2011]. The assumptions in univariate analysis models are often an over-simplification of dynamic PET data, as region-specific binding is not independent between regions, and may be both functionally and structurally connected [Beliveau et al., 2015]. However, univariate analysis models are simple and provide a straight-forward interpretation of a single variable.

Multivariate analysis models account for the correlation/covariance structure between brain regions, identifying spatially distributed patterns of specific binding that fluctuate coherently with group, session and/or other external variables (Figure 2.7). The application of multivariate analysis models can range from simple linear models (approaching univariate techniques) to complex non-linear models [Hansen et al., 1999, Morch et al., 1997]. Common for all multivariate analysis models is that each set of N brain regions is treated as an N-dimensional vector, and each scan is treated as a set of N-dimensional data points. The multivariate analysis model then searches for a lower-dimensional vector (discriminant) that best discriminates spatial patterns of brain regions that are different between conditions (e.g. test vs. retest). Multivariate analysis models are sensitive to cases where brain regions have strong spatial correlations, a condition that is satisfied in PET data. Taking this set of co-varying brain regions, there exists a linear combination that can capture a more sensitive signal (Figure 2.7)

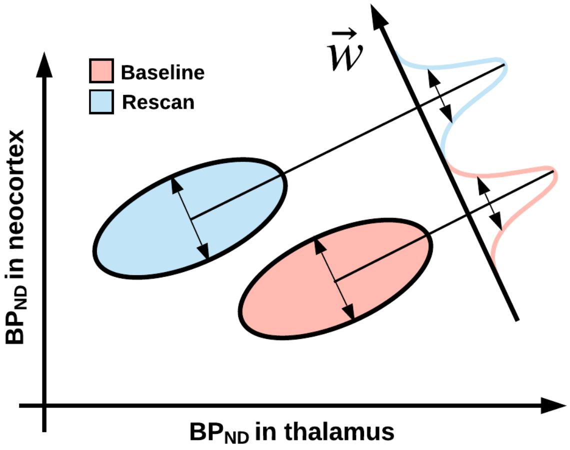

Figure 2.7: Multivariate normal distributions of BP_{ND} in the thalamus (first axis) and in the neocortex (second axis) across baseline (red distribution) and rescan (blue distribution). If measured univariately (i.e. either on the first or second axis) the distributions are not separable, but there exists a projection onto the vector $\vec{w}$ that separates the two distributions.

Both univariate and multivariate analysis models have strengths and weaknesses. Although univariate analysis models have been widely employed in the PET community, the existence of complex dependencies between brain regions may not be fully explained by univariate models, biasing the model at the expense of decreased variance [Nørgaard et al., 2017]. Biased univariate analysis models therefore tend to be more robust, reproducible and less sensitive to preprocessing strategies compared to less-biased multivariate models that have higher variance. In comparison, multivariate analysis models are more sensitive to weaker and spatially distributed patterns of signal, and may have better detection of signal if preprocessed optimally. In this sense there is a bias-variance trade-off to consider. However, while univariate models need to correct for multiplicities due to multiple hypothesis testing, multivariate analysis models allow for alternative significance tests that do no require correction for multiple comparisons. The different approaches for significance testing have different assumptions, and it is therefore likely that they have different control over the probability that the positive conclusions could arise under the null hypothesis (false discovery rate, FDR). Objectively, it is difficult to argue, that multivariate analysis models perform better than univariate analysis models, as these methods, in fact, provide different inferences about the data. In the functional MRI (fMRI) literature, it has been shown that the distinction between univariate and multivariate analysis models is dependent on the preprocessing, with better preprocessing strategies minimizing the distinction [Tegeler et al., 1999, Churchill et al., 2015]. Nevertheless, **there is a need in the PET community to examine the interaction between statistical analysis and preprocessing optimization, and its influence on the false-positive rate**. Throughout the thesis, I apply the multivariate Linear Discriminant Analysis (LDA) model for prediction of two-class classification problems. The LDA model is sensitive to preprocessing choices, and is conceptually an advantageous approach in cases where there exists an a priori hypothesis that differences in $\mathrm{BP}_{\mathrm{ND}}$ are not regional, but rather occur on a network level with spatially distributed patterns of $\mathrm{BP}_{\mathrm{ND}}$.

2.3 Optimization of Preprocessing Strategies

In the absence of a “ground truth”, it remains a major challenge in PET to optimize preprocessing strategies, and it may take alternative performance metrics to quantitatively evaluate and compare various preprocessing strategies [Strother et al., 2002, Churchill et al., 2015]. The uptake of radioligand in the brain varies across both regions and subjects, but also between scan sessions [Frankle et al., 2004]. Therefore, there exists no unifying pattern of radioligand uptake that can be predicted and generalized to the population. Simulations can overcome these latter limitations of real data by providing the "ground truth", having knowledge about the true underlying data generating process [Ichise et al., 2003]. However, while simulations can be instructive, it is obviously very difficult to simulate the complex spatio-temporal noise patterns arising from a PET scanner. Simulations therefore provide only limited information on preprocessing effects. Two broad categories of performance metrics will be used in this thesis to quantitatively measure pipeline effects in real data: (1) Reproducibility and (2) Prediction. The reproducibility metrics are computed using statistical subsampling, and the prediction metric is computed using nested cross-validation. The rationale for these metrics are listed below.

2.3.1 Performance Metrics

Reproducibility: To capture the variation of $\mathrm{BP}_{\mathrm{ND}}$ in different brain regions, between subjects and between sessions, new PET radioligands are typically examined in a test-retest setting, implicitly assuming that test and retest should generate similar outcomes [Ogden et al., 2007, Kim et al., 2006]. It seems to be the understanding that the test-retest examination is the ultimate validation for successful application of the radioligand in the community. For example, if the between-subject variation is too high it may require an unreasonable number of subjects to establish group differences. Furthermore, if the within-subject variation is too high it becomes infeasible to perform longitudinal studies applying a pharmacological intervention, e.g. if the expected within-subject variability is larger than the effect of the intervention. For these reasons, performance metrics of reproducibility such as test-retest bias, within- and between-subject variance, and the Intraclass-Correlation Coefficient (ICC) are often applied in PET test-retest studies [Frankle et al., 2004, Ogden et al., 2007, Kim et al., 2006]. The details of these metrics are provided in Chapter 3.

Prediction: Models providing a prediction metric (e.g. predictive accuracy) are conceptually intriguing compared to univariate hypothesis testing, as they provide a quantitative measure of the ability to correctly predict the experimental condition (class) in an independent sample [Varoquaux et al., 2017]. Specifically, a predictive model is built on a training data set to estimate the discriminant brain pattern that dissociates between classes. Subsequently, the model is evaluated on its ability to predict the classes in an independent data set.

Although the goal of any neuroscientific study is to maximize model prediction and reproducibility, these metrics represent unique trade-offs in model parameterization, usually limiting the ability to maximize both simultaneously [Baldassarre et al., 2017]. For example, models driven by maximization of reproducibility will have stable and reproducible brain patterns, but will have less sensitivity towards detecting minor changes in binding following a pharmacological intervention. An illustration of this can be made by considering the application of a infinitely high smoothing kernel to the PET data. The output will be a perfectly reproducible brain pattern, but the analysis model will have no ability to predict the experimental condition due to the lack of variation. In contrast, models maximizing prediction will be highly sensitive towards predicting minor changes in binding, but will tend to produce non-reproducible and unstable brain patterns. In this thesis, the performance of a given preprocessing strategy is first evaluated using metrics of reproducibility. Then, preprocessing strategies are evaluated for their predictive performance in an independent test set. Finally, a joint evaluation is carried out to identify a compromise between model parameterizations of prediction and reproducibility, which ultimately can be used to select an optimal preprocessing strategy.

2.4 Thesis Objectives

The overall goal of this thesis is to improve signal detection in dynamic PET imaging studies by evaluating and optimizing choices in the preprocessing pipeline, using statistical performance metrics of reproducibility and prediction.

Goal 1: To identify the variability of acquisition and preprocessing choices in the [^{11}C]DASB-PET literature, and to quantify the impact of the choices using a meta-analytic approach. I will review data from 21 PET centres that published a total of 105 [^{11}C]DASB-PET papers between November 2000 and March 2017 (manuscript A).

Goal 2: To evaluate the impact of commonly used PET preprocessing strategies (addressed in **Goal 1**) on a test-retest data set.

> Goal 2a: I will examine 384 different strategies in 30 subjects that were scanned twice with the 5-HTT radioligand [^{11}C]DASB. Five commonly used preprocessing steps, each with 2-4 plausible options, will be investigated: (1) motion correction (MC), (2) co-registration, (3) delineation of volumes of interest (VOI's), (4) partial volume correction (PVC), and (5) kinetic modeling (manuscript B).
>
> Goal 2b: Examine the impact of preprocessing strategies on the false-positive rate in univariate and multivariate analysis models with and without correction for multiple comparisons (manuscript C).

Goal 3: Examine the impact of pipeline choices, as indexed in Goal 1 and Goal 2, in an independent test-set of 30 subjects with a pharmacological intervention between scan 1 and scan 2 (manuscript D).

Goal 4: Develop a statistical framework for estimation of statistical significance in the context of multiple preprocessing strategies and predictive classification (manuscript E).

For each **Goal**, I focus on the characterization of preprocessing optimization, and discuss how each performance metric provides valuable information. In Chapter 3, I provide the methodological aspects of the PET workflow that has been used in this thesis, ranging from subject selection, data acquisition, preprocessing and statistical analysis (Figure 2.4). In Chapter 4, I address **Goal 1** by evaluating the variety of methodological choices in the [^{11}C]DASB-PET literature. In Chapter 5, I examine a subset of the choices identified in **Goal 1** to address **Goal 2**: the effects of commonly used PET preprocessing strategies on performance metrics of reproducibility (**Goal 2a**) and false-positive rate (**Goal 2b**) in a test-retest data set. Chapter 6 expands on the results from **Goal 1** and **Goal 2**, by determining the extent to which different preprocessing strategies lead to different conclusions (**Goal 3**) in a randomized, double-blind, placebo-controlled study using a pharmacological intervention. Chapter 7 establishes a non-parametric 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 (**Goal 4**). Finally, Chapter 8 summarizes the outcomes from **Goal 1-4** and discusses future research objectives.

Chapter 3

Methods - The Data-Analysis Chain

In this chapter, I will review the details of the Data-Analysis Chain that is used throughout **Chapters 5-7** in this thesis.

3.1 Subjects

A total of 60 female participants (mean age 24.3 ± 4.9 years) were included in a double blind, randomized, placebo-controlled study [Frokjaer et al., 2015] investigating depressive responses to sex-steroid hormone manipulation. Participants received either a subcutaneouos injection of a gonadotropin releasing hormone agonist (GnRHa) implant (ZOLADEX with 3.6 mg of goserelin; Astra Zeneca, London, UK) (N=30) or saline (N=30). All participants were PET scanned twice on separate days (median interval of 34 days). One subject in the GnRHa group was excluded due to an issue with the PET acquisition, leaving 29 subjects available for analysis. Further details 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. The 30 subjects receiving placebo were used in manuscripts [B] and [C], whereas the remaining 29 subjects receiving an intervention were used in manuscripts [D] and [E]. The placebo group was considered to represent test-retest conditions with no expected changes between scan 1 and scan 2. I therefore used this data set to optimize the preprocessing strategy using a set of performance metrics related to reproducibility. The remaining part of the data including the active intervention was used to optimize the preprocessing strategy using predictive accuracy as performance metric. The details of the evaluation and optimization are provided below.

3.2 Data Acquisition

In this thesis, a fixed dynamic PET sequence with parameter settings commonly used in the literature, was employed using the Siemens ECAT High-Resolution Research Tomography PET scanner. The data was acquired in 3D list-mode and with the highly selective radioligand [^{11}C]DASB. The imaging protocol consisted of a single-bed, 90 minutes emission acquisition post injection of 587 ± 30 (mean ± 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 ([Olesen et al., 2009]) 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]).

3.3 Preprocessing

Here, I establish a 5-step data preprocessing pipeline, each step with 2 to 4 options, to estimate the outcome measure $\mathrm{BP}_{\mathrm{ND}}$ (Figure 3.1). All the individual procedures have previously been used in published [^{11}C]DASB-PET studies, except for PVC using the GTM. The steps are listed below in the order in which they were applied. Specific rationales for including/excluding each unique preprocessing step and their options are listed below.

Step 1 – Motion Correction (with/without)
Within-scan PET motion correction was executed using a data-driven automated image registration (AIR v. 5.2.5, http://loni.usc.edu/Software/AIR). Prior to alignment, each frame was smoothed using a 10 mm Gaussian 3D kernel and thresholded at the 20-percentile level to boost SNR. Alignment parameters were estimated for the smoothed PET frames 10-36 to a reference frame with high SNR (frame 26) using a scaled least squares cost-function in AIR. Subsequently, the non-smoothed frames were transformed using the estimated alignment parameters and resliced into a 4D motion corrected data set (e.g., as applied in [Frokjaer et al., 2015] and [Beliveau et al., 2017]). The motion correction estimation for frame 10 was applied to the first 9 frames. I chose to register frames 10-36 only, because the first 9 time frames (10/20 sec) have low count statistics, high noise levels and have shown to produce highly variable alignment parameters. Criterion for acceptable motion was a median movement less than 3 mm across frames, as estimated by the median of the sum of the squared translations (x,y,z) across all voxels. The rationale for testing the effect of MC in the pipeline is because motion artefacts vary by dataset. Furthermore, MC should ultimately control motion artefacts, but may also impose unwanted biases on the data or reduce experimental power, especially in cases of minor or no head movement [Churchill et al., 2012]. In addition, manuscript [A] showed that MC lowers be-

tween subject variability in striatum, resulting in 26% fewer subjects needed in a group analysis to achieve similarly power statistical tests. It is therefore of interest to validate this latter observation in an independent data set.

Step 2 – Co-registration (4 options)
All single-subject PET frames were initially either summed (according to their frame length i.e. integral) 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 and Fischl, 2009]) each with different cost functions. This step is explicitly evaluated, as its effects may vary by dataset and as a function of SNR.

Step 3 – Delineation of Volumes of Interest (3 options)
MRI scans were processed using FreeSurfer (http://surfer.nmr.mgh.harvard.edu, version 5.3). FreeSurfer contains a fully automatic structural imaging pipeline for processing of cross-sectional as well as longitudinal data. Furthermore, it includes several features such as skull stripping, B1 bias field correction, non-linear registration to a stereotaxic atlas, statistical analysis of morphometric differences, and probabilistic labeling of cortical/subcortical brain structures based on the Desikan-Killiany atlas [Fischl et al., 2004]. A total of 28 subcortical and cortical regions were extracted, and averaged across hemispheres producing a final sample of 14 regions pr. subject/pipeline. The volumetric regions included the amygdala, thalamus, putamen, caudate, anterior cingulate cortex (ACC), hippocampus, orbital frontal cortex, superior frontal cortex, occipital cortex, superior temporal gyrus, insula, inferior temporal gyrus, parietal cortex, and entorhinal cortex. These regions were chosen because they largely cover the entire brain, but also because many of the regions have been used in previously published DASB PET studies. Out of more than 100 published [^{11}C]DASB-PET studies [A], each region is mentioned N times: amygdala (N=72), thalamus (N=105), putamen (N=88), caudate (N=82), ACC (N=74), hippocampus (N=71), frontal cortex (N=66), occipital cortex (N=48), temporal cortex (N=58), parietal cortex (N=34), entorhinal cortex (N=16). Subsequently to running the FreeSurfer pipeline, the user can choose to perform user-dependent manual edits to the FreeSurfer output, to correct for errors mostly located in the white matter (WM), cerebrospinal fluid (CSF) or on the pial surface. The manual editing was carried out according to FreeSurfer recommendations (https://surfer.nmr.mgh.harvard.edu). If a T2-weighted MRI is also available, semi user-independent edits can also be made to the FreeSurfer output by re-running the FreeSurfer reconstruction including the T2-weighted MRI. I examined all three pipelines in this thesis and now refer to these as FS-RAW (standard output from FreeSurfer), FS-MAN (output from FreeSurfer with manual edits) and FS-T2P (output from FreeSurfer with the T2 stream). Only the first test-scan MRI was used for the analysis. Different FreeSurfer options are tested, as the optimal correction for errors has been reported to vary as a function of subject and scanner [McCarthy et al., 2015].

Although choice of atlas (e.g. PVElab, AAL or MNI305) may have an impact on the outcome, I considered assessment of various atlas choices to be beyond the scope of the current work and I consistently applied the Desikan-Killiany atlas provided in FreeSurfer.

Step 4 – Partial Volume Correction (4 options)

The data were analyzed either without or with three PVC approaches. The VOI-based PVC technique, Geometric Transfer Matrix (GTM), by [Rousset et al., 1998] was applied using PETsurfer (https://surfer.nmr.mgh.harvard.edu/), by for each frame (1) establishing a forward linear model relating [^{11}C]DASB uptake values to the VOI means (Equation 3.1) and then (2) solving it using the inverse Equation 3.2.

$$y = \mathbf{X}\beta \tag{3.1}$$

$$\hat{\beta} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T y \tag{3.2}$$

where y is a column vector with elements of [^{11}C]DASB uptake values and with length corresponding to the number of voxels, $\mathbf{X}$ is a design matrix with size equal to the number of voxels (rows) and number of VOIs (columns), β is a row vector with length equal to the number of VOIs and represents the true underlying VOI means, and $\hat{\beta}$ is the estimate of the VOI means. The design matrix $\mathbf{X}$ was computed by (1) for each VOI, a sparse image of tissue fraction[1] (TF) values for that VOI was created in PET space (2) this image was then smoothed with a Gaussian kernel corresponding to the PSF of the scanner, and (3) the image was then reshaped into a column vector and stored in the corresponding column in $\mathbf{X}$. This procedure was repeated for all VOIs.

Because the PSF for a HRRT scanner reconstructed with a OP-OSEM-PSF algorithm varies from 1-2.5 mm in radial orientation depending on the distance from the centre of the field of view ([Olesen et al., 2009]), I ran the analyses with the PSF settings; 0 mm and 2 mm. However, because motion, inhomogeneous tracer uptake and varying uptake across frames is likely to further increase the spatial resolution as compared to a point source in [Olesen et al., 2009], I also ran the PVC analyses with a 4 mm PSF, as used in [Greve et al., 2014]. The PVC step is evaluated, because it has been suggested to be the optimal solution for VOI analysis, given that assumptions about the PSF, accurate delineation of regions, correct PET-MRI registration, and constant uptake within each VOI are satisfied ([Greve et al., 2016]). In addition, a homogeneous CSF and WM segmentation is

---

[1] The TF effect is the result of a voxel occupying multiple tissue types. The TF is the fraction of tissue type (i.e. GM, WM, CSF) inside a given voxel. To create a segmentation in PET space, each voxel was assigned to the VOI with the highest TF value in that voxel.

important (provided in FreeSurfer), as these are primary regions to compensate for in gray matter uptake of the tracer. When the assumptions are satisfied (and under noiseless conditions), the GTM will provide the exact mean in each VOI.

Step 5 – Kinetic Modeling (4 options)

The Multilinear Reference Tissue Model (MRTM) was applied as described by [Ichise et al., 2003] with cerebellum (excluding vermis) as a reference region, allowing for estimation of three parameters from which the $\mathrm{BP_{ND}}$ can be derived. The operational equation for MRTM (Equation 3.3) is formulated as,

$$C(T) = -\frac{V}{V'b}\int_0^T C^{'}(t)dt + \frac{1}{b}\int_0^T C(t)dt - \frac{V}{V'k_2^{'}b}C^{'}(T) \tag{3.3}$$

where $C(t)$ is the radioligand distribution (MBq/mL) in the target region at time t, $C^{'}(t)$ is the radioligand distribution (MBq/mL) in the reference region, V and $V^{'}$ are the corresponding total distribution volumes (mL/mL), $k_2^{'}$ is the transfer from reference to plasma (min^{-1}), and b is the intercept term.

The second model applied was the Multilinear Reference Tissue Model 2 (MRTM2) ([Ichise et al., 2003]) with cerebellum (excluding vermis) as a reference region (Equation 3.4). Thalamus, putamen and caudate were averaged to represent a single less noisy high-binding region for estimation of the rate constant, $k_2^{'}$, using the MRTM model from Equation 3.3,

$$C(T) = -\frac{V}{V'b}\left(\int_0^T C^{'}(t)dt + \frac{1}{k_2^{'}}C^{'}(T)\right) + \frac{1}{b}\int_0^T C(t)dt \tag{3.4}$$

The MRTM2 is similar to MRTM, except that $k_2^{'}$ is determined after an initial iteration of MRTM and its value is subsequently entered into the two-parameter MRTM2 model. This approximates to a linear kinetic analysis, but is executed in only a fraction of the computational time.

The third model applied was the simplified reference tissue model, SRTM, as described by [Lammertsma and Hume, 1996] (Equation 3.5).

$$C(t) = R_1C^{'}(t) + \left(k_2 - \frac{R_1k_2}{(1+BP_{ND})}\right)C^{'}(t) \otimes e^{\frac{-k_2t}{(1+BP_{ND})}} \tag{3.5}$$

SRTM allows for nonlinear least squares estimation of three parameters (R_1, k_2 and $\mathrm{BP_{ND}}$) from each TAC. R_1 is the relative radioligand delivery and k_2 is the rate constant for transfer from free to plasma (min^{-1}).

The non-invasive Logan reference tissue model was applied as described in [Logan et al., 1996] with t* = 35 minutes for all regions and subjects (Equation 3.6)

$$\frac{\int_0^T C(t)dt}{C(T)} = DVR\left[\frac{\int_0^t C'(t)dt + \frac{C'(T)}{k_2'}}{C(T)}\right] + b \tag{3.6}$$

where DVR is the distribution volume ratio. The non-invasive Logan also assumes the existence of a valid reference region and requires the estimation of k_2', similarly as for MRTM2. The $\mathrm{BP_{ND}}$ can subsequently be estimated as $BP_{ND} = DVR - 1$. All kinetic models applied in this work were implemented in MATLAB v. 2016b as specified in their original paper. The implementation in MATLAB was validated with PMOD v. 3.0 (10 subjects $< 0.1\%$ difference in $\mathrm{BP_{ND}}$), but was carried out in MATLAB for parallel execution purposes to substantially reduce processing time. Different kinetic modeling approaches are tested in this thesis, as the optimal estimation of 5-HTT binding may vary as a function of SNR, subject and region.

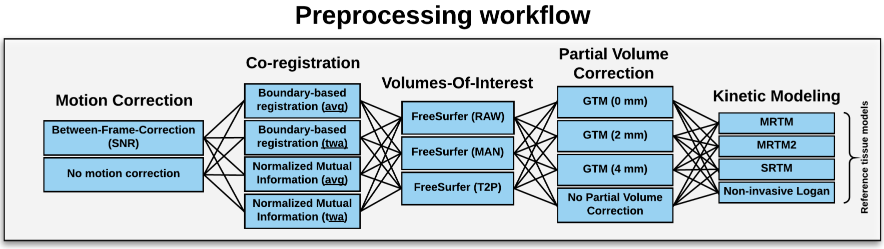

Figure 3.1: Schematic overview of the various preprocessing steps applied for the [^{11}C]DASB quantification. There are 384 different preprocessing strategies in total. Abbreviations; average (avg), time-weighted average (twa), signal-to-noise ratio (SNR), Geometric Transfer Matrix (GTM).

3.4 Statistical Analysis

3.4.1 Performance Metrics of Reproducibility

In this section, I introduce the performance metrics of reproducibility used to measure the effect of different preprocessing choices in the test-retest data consisting of 30 participants. While most of these metrics were computed for each region k and summarized over subjects i, I also adopted a reproducibility metric from the fMRI literature producing a single reproducibility measure for each subject i and pipeline j, using the linear relationship between all VOIs [Strother et al., 2002]. I used statistical subsampling to evaluate sample sizes of either $\tilde{n} = 10$ or 20 subjects randomly selected without replacement from the 30 subjects, and this was repeated 1000 times to compute a mean estimate and a 95% confidence interval (CI). Notation-wise, $\tilde{n}$ indicates a resampling analysis, whereas N = 30 indicates that all subjects were used to compute the metric. Statistical differences in pipeline choice (e.g., motion correction vs. no motion correction (nMC)) for each performance metric was determined across 1000 resamples (subsampling 20 subjects without replacement), and then using the empirical distribution of the differences of the performance metric. This provides an empirical p-value for the difference between pipeline choices for each performance metric. Correction for multiple comparisons across regions was carried out using False-Discovery Rate (FDR, [Benjamini and Hochberg, 1995])), at FDR=0.05. $\mathrm{BP_{ND}}$'s that were less than 0 or larger than 10 in either test or retest were excluded in all estimations to avoid the influence of outliers.

Test-retest bias
The test-retest bias was computed as the difference between the two measurements and expressed as a percentage relative to the first scan,

$$Bias_{i,j,k} = 100 \times \frac{retest_{i,j,k} - test_{i,j,k}}{test_{i,j,k}} \tag{3.7}$$

Within-Subject Variability
The within-subject variability was computed as the standard deviation of the bias [Kim et al., 2006], and then normalized to a coefficient of variation (expressed in percent) by dividing by the group average value, μ,

$$WSV_{j,k} = 100 \times \left(\frac{\sqrt{\frac{\sum_{i=1}^{\tilde{n}} (d_{i,j,k} - \bar{d}_{j,k})^2}{n-1}}}{\frac{\sum_{i=1}^{S} (test_{i,j,k} + retest_{i,j,k})/2}{S}} \right) \tag{3.8}$$

where $d_{i,j,k} = test_{i,j,k} - retest_{i,j,k}$, $\bar{d}_{j,k} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} d_{i,j,k}$ and S is the number of sub-samples (i.e. outliers excluded).

Between-Subject Variability

The between-subject variability was computed as the between-subject standard deviation, σ, and then normalized to a coefficient of variation (expressed in percent) by dividing by the group average, μ,

$$BSV_{j,k} = 100 \times \frac{\sigma_{j,k}}{\mu_{j,k}} \tag{3.9}$$

Intra-Class Correlation

The test-retest reliability was estimated using the intra-class correlation coefficient (ICC),

$$ICC_{j,k} = \frac{MSBS_{j,k} - MSE_{j,k}}{MSBS_{j,k} + (q-1)MSE_{j,k}} \tag{3.10}$$

where MSBS is the mean sum of squares between subjects, MSE is the mean squared error, and q is the number of within-subject measurements (= 2 in our case).

Sample Size Estimation

The needed sample size, $\hat{n}$, to show an effect E at a 95% confidence level for pipeline j and region k was computed as,

$$\hat{n}_{j,k} = \left(\frac{1.96 \times \sigma_{j,k}}{E_{j,k}} \right)^2 \tag{3.11}$$

where σ is the BSV.

Global Signal-To-Noise Ratio (gSNR)

A global reproducibility metric (gSNR) was computed for each subject i and pipeline j, by taking the pairwise linear correlation based on the Pearson correlation coefficient (R) over all regions based on test and retest $\mathrm{BP}_{\mathrm{ND}}$'s,

$$gSNR_{i,j} = \sqrt{\frac{(1 + R_{i,j}) - (1 - R_{i,j})}{(1 - R_{i,j})}} \tag{3.12}$$

Initially described in [Churchill et al., 2012], the optimal fixed pipeline across regions was computed as: (1) for each subject, rank the pipelines 1-384 based on gSNR, with a higher rank indicating better performance (i.e., higher gSNR). Then (2) compute the median ranking across subjects, and select the pipeline with highest median rank as the optimal fixed choice (FIX), and (3) perform a non-parametric Friedman test on the pipeline rankings to determine if there is a significant ordering of fixed pipelines across subjects. If this test is significant, perform a post-hoc estimation of the critical-difference cut-off at α=0.05, based on the Dunn-Sidak test. If the difference in median ranks between two pipelines is higher than the cut-off, it indicates that the pipelines are not statistically distinguishable in performance. This procedure may be used to identify a list of optimal fixed pipelines that are not significantly worse than FIX.

3.4.2 Statistical Analysis Models

The main **univariate analysis model** applied in this thesis was the paired t-test, determining whether the mean difference in $\mathrm{BP_{ND}}$ between two sessions as a function of pipeline j and region k is zero. All data was tested for normality using a Kolmogorov-Smirnov (KS) test. Within each pipeline, j, the regions were corrected for multiple comparisons using FDR at q = 0.05. A p-value less than 0.05 was considered a significant result.

Throughout this thesis, I also deployed a **multivariate Linear Discriminant Analysis (LDA) model** for prediction of two-class classification problems. For this two-class dataset $\mathbf{X}$, 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:

$$p(\mathbf{X}|C_k;\theta) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}||\mathbf{L_{train}}^T(\mathbf{X} - \bar{\mathbf{X}}^k_{\mathbf{train}})||^2\} \tag{3.13}$$

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 Equation 3.13, 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 in a 5-fold cross-validation framework (Figure 6.1). The model was then evaluated using a validation set, $\mathbf{X}$, consisting of the remaining 20%. 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 labels 1000 times and rerunning the above 10 randomized 5-fold cross-validation procedure to generate an empirical null-distribution. This provides an empirical p-value for each model and pipeline.

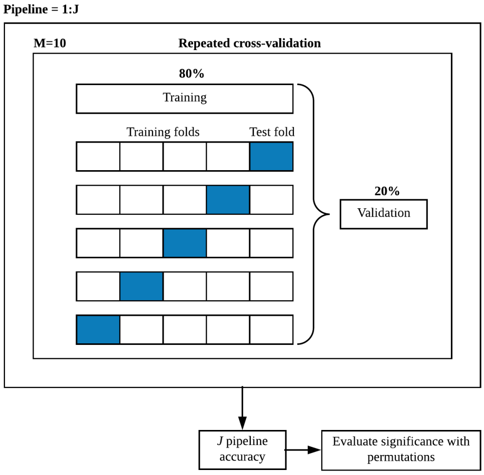

Figure 3.2: Overview of the nested cross-validation schema with M repetitions, 80% training data, and 20% validation data, for each pipeline j.

Chapter 4

Study 1: Preprocessing Strategies in the PET Literature

4.1 Introduction

This chapter is adapted from the peer-reviewed journal article [A]: Nørgaard M, Ganz M, Svarer C, Feng L, Ichise M, Lanzenberger R, Lubberink M, Parsey RV, Politis M, Rabiner EA, Slifstein M, Sossi V, Suhara T, Talbot PS, Turkheimer F, Strother SC, Knudsen GM. 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, 2019 Feb;39(2):210-222. DOI: 10.1177/0271678X18770107.

Since the introduction of [^{11}C]DASB-PET in November 2000 [Houle et al., 2000], hundreds of PET studies have been carried out. Several of these studies have reported less 5-HTT binding in depressed patients compared to healthy individuals [Parsey et al., 2006a, Hammoud et al., 2010]. Other studies have reported on the relationship between levels of 5-HTT occupancy and doses of SSRIs to achieve a therapeutic effect [Meyer et al., 2001, Parsey et al., 2006b]. Common for all [^{11}C]DASB-PET studies is that they rely on valid quantification of 5-HTT binding to produce valid results and conclusions. However, it is currently unclear how much the data-analysis chain varies in the literature, and more importantly how these variations may have affected the reported findings. In this Chapter, I systematically review the literature on differences in subject selection, data acquisition and preprocessing in 105 studies applying the radioligand [^{11}C]DASB. To quantify the influence of each step, I also extract the available average $\mathrm{BP}_{\mathrm{ND}}$'s in healthy participants in the striatum and ACC from 90 of the 105 studies, and use linear models to associate the BP_{ND}'s with the different steps in the data acquisition and preprocessing stages. To ensure accuracy in my interpretations, I invited relevant co-authors of the 105 published studies to make contributions and comment on the work.

4.2 Methods

In this section, I establish the procedures for generating an overview of the available steps of data acquisition and preprocessing in $[^{11}C]$DASB-PET studies in the literature.

4.2.1 The Data-Analysis Chain

Subject Selection: The variation in subject selection across studies was extracted by categorizing each group of subjects into either healthy participants (control group) or a specific patient cohort (e.g. Alzheimer's Disease).

Data Acquisition: The variation in data acquisition across studies was established by defining six main categories that could vary between PET centres, scanners and subjects. These categories included the (1) MRI scanner type, (2) PET scanner type, (3) PET scan duration, (4) number of frames, (5) injected dose, and (6) reconstruction type.

Preprocessing: The variation in preprocessing across studies was extracted by categorizing each study into all five main categories (1) MC, (2) co-registration, (3) VOI technique, (4) PVC technique, and (5) kinetic modeling.

Statistics: The available group average BP_{ND} and standard deviation in the striatum and ACC was extracted for healthy participants from each study. These values were used as the dependent variable in separate linear regression models, with the independent variables: number of participants in the study, age, age standard deviation, MRI scanner type, PET scanner type, number of frames, injected dose, MC (yes/no), VOI technique, and kinetic modeling technique. All covariates were standardized columnwise to have mean 0 and standard deviation 1. There were 50 studies reporting striatal BP_{ND} and 43 studies reporting ACC BP_{ND} that had information for all of the independent variables.

4.3 Results

4.3.1 Subject Selection

Figure 4.1 shows the number of available $[^{11}C]$DASB-PET data sets in the literature and as a function of time. The majority of $[^{11}C]$DASB-PET data sets are healthy subjects, summing to a total of 1856 available data sets. The second largest group is Major Depressive Disorder (MDD) with 234 available data sets. In a linear model with between-subject variation in ACC as the dependent variable, and the independent variables (i.e. number of participants in the study, age, age standard deviation, MRI scanner type, PET scanner type, number of frames, injected dose, MC (yes/no), VOI technique, and kinetic modeling technique), I identified a significant trend for an age effect (p = .075, uncorrected), suggesting that BP_{ND} is more variable in elderly than in younger subjects.

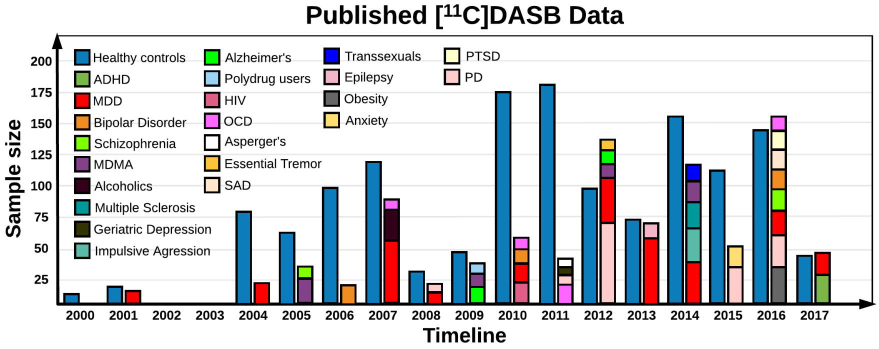

Figure 4.1: Timeline of number of patient and healthy controls in the 105 published $[^{11}C]$DASB studies. The colors indicate either healthy controls, or a specific disorder as a function of time and sample size. ADHD: attention-deficit/hyperactive disorder; MDD: major depressive disorder; MDMA: ecstasy; HIV: human immunodeficiency virus; OCD: obsessive compulsive disorder; SAD: seasonal affective disorder; PTSD: post-traumatic stress syndrome; PD: Parkinson's disease.

4.3.2 Data Acquisition

Across 21 PET centres, 9 different PET scanners have been used (Figure 4.2). The most commonly used PET scanners are the ECAT EXACT HR+ scanner and the GE Advance scanner, having a spatial resolution ranging from 4.3-8.3 mm FWHM. The MRI scanners vary in field strength from 0.3T to 7.0T, with the most widely used acquisitions of 1.5T (43%) or 3.0T (32%). The duration of the dynamic PET scan varied from 30-120min (30,60,80,90,95,100,110,120min), with 90 minutes being the most frequent choice. The framing of the dynamic PET data had been used in 17 different ways, ranging between 17 and 50 frames, with 26 frames being the most common. The injected dose varied from approximately 100 MBq to 740 MBq across subjects and studies. Finally, the main types of reconstruction that had been applied were either Filtered backprojection (FBP) or Ordered-Subset Expectation Maximation (OSEM), with FBP being the most frequent. To summarize the data acquisition stage, the most widely published workflow consists of: 1.5T MRI (43%), ECAT EXACT HR+(43%), 90-min acquisition (65%), 26 frames (17%), and FBP to reconstruct the 4D PET data (72%).

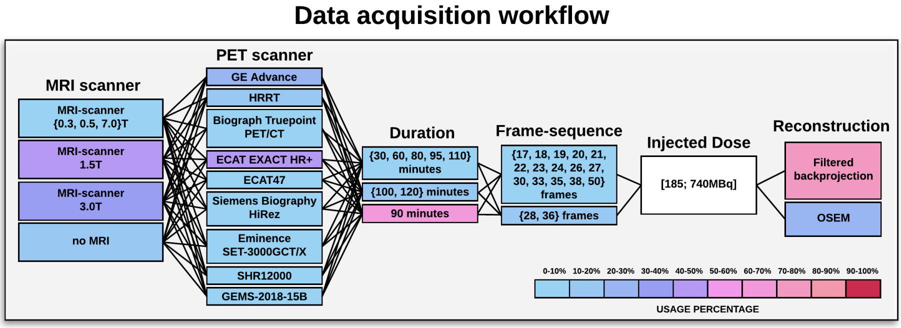

Figure 4.2: Schematic overview of the different data acquisition workflows used to acquire dynamic [^{11}C]DASB data. The workflow consists of scanners providing anatomical information, i.e. MRI scanners at various field strengths (Tesla), various PET scanners, duration of the dynamic PET acquisition, frame sequence used to temporally acquire 4D [^{11}C]DASB data, injected dose (ranging from approximately 100-740 MBq), and finally the reconstruction methods used to reconstruct the 4D PET sequence. The colors indicate the frequency per step that has been applied in a [^{11}C]DASB PET study out of the total 105 studies. Injected dose is filled as white, because it spans a continuous range and is highly subject-specific. The 4D imaging data are the output of the data acquisition workflow and input to the preprocessing workflow.

4.3.3 Preprocessing

Between-frame MC of the PET data was applied in 59% of studies, whereas 41% left out MC. The MC procedure mainly varied between the registration to either (1) a frame with high SNR or (2) a mean/summed PET image. Co-registration between PET and MRI was mainly carried out using NMI (98%), whereas the remaining 2% used BBR. The PET image used for co-registration was predominantly either a frame with high SNR or a mean/summed image. For delineating VOIs, 8 different techniques had been applied, with manual delineations being the most frequent (38%). PVC had only been applied in 4 published studies. Kinetic modeling was applied in 9 different ways, mainly dividing the methods into reference tissue methods and methods using an arterial input. The most commonly applied kinetic model was the MRTM2 (38%). In a linear model with between-subject variation in striatum as the dependent variable, and the independent variables (i.e. number of participants in the study, age, age standard deviation, MRI scanner type, PET scanner type, number of frames, injected dose, MC (yes/no), VOI technique, and kinetic modeling technique), I identified a significant trend for an effect of MC (p = .064, uncorrected), suggesting that MC lowers between-subject variability with 0.035 compared to data without MC. This translates into 26% fewer subjects needed in a group analysis to obtain the same statistical power.

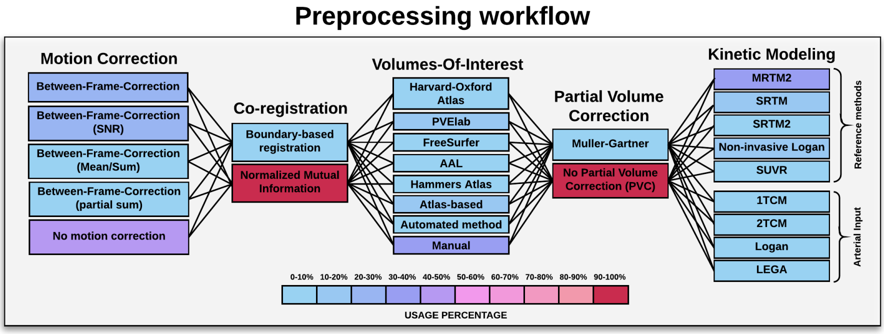

Figure 4.3: Schematic overview of the various preprocessing steps used in analyzing dynamic [^{11}C]DASB data. This ranges from different motion correction techniques, co-registration, volume-of-interest definitions, partial volume correction, and kinetic modeling. The colors indicate the percentage, in which a given step has been applied in the 105 [^{11}C]DASB-PET studies.

4.3.4 Statistical Analysis

Figure 4.4 shows a histogram of the group average BP_{ND} and between-subject variability (expressed as a coefficient of variation, CV) in the striatum and across studies. The CV ranged from 2.8% to 24.4% with the majority of studies producing a CV ranging from 3 to 11%. The sample size varied from 4 to 144 with the majority of studies using between 10 and 20 subjects.

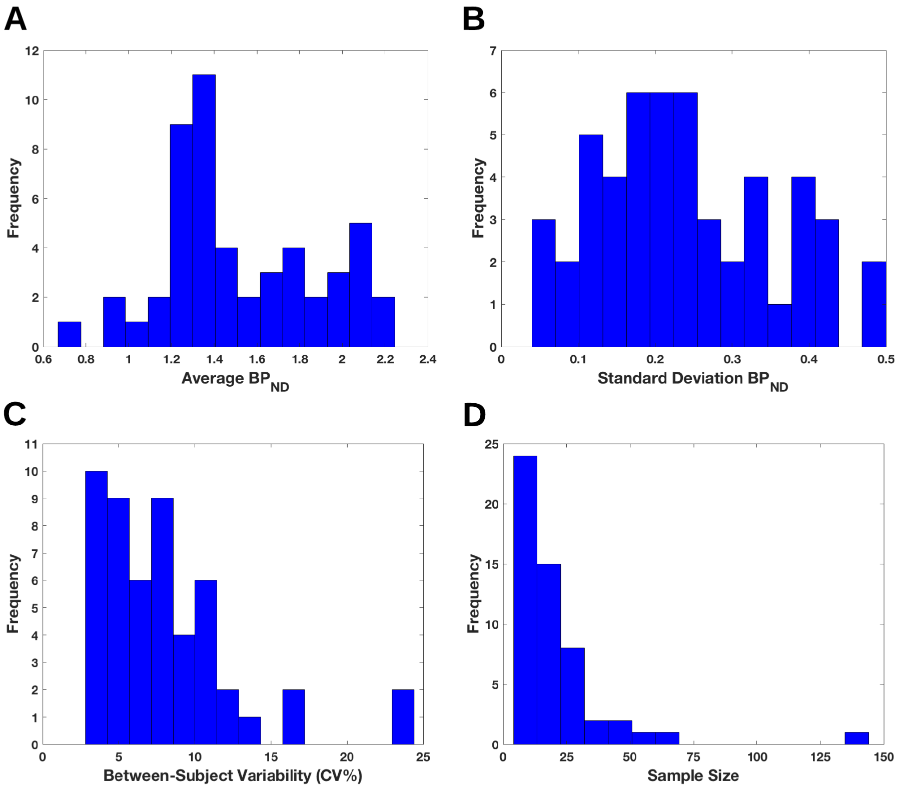

Figure 4.4: Striatal **(A)** group average $\mathrm{BP}_{\mathrm{ND}}$ **(B)** standard deviation **(C)** between-subject variability expressed as a coefficient of variation (ratio of the standard deviation to the mean, CV), and **(D)** sample size in groups of healthy participants across 50 [^{11}C]DASB-PET studies.

4.4 Discussion

In this Chapter, I demonstrated that most [^{11}C]DASB-PET experiments are performed under the implicit assumption that the results they generate are either (1) insensitive to the preprocessing strategy or (2) standard preprocessing strategies produce near-optimal results. Combinatorially, there are 21.150.720 different workflows in the [^{11}C]DASB-PET literature that have been used for quantification of 5-HTT binding. For preprocessing only, there are at least 1440 combinations that have been applied in the literature, ranging from differences in MC, co-registration, delineation of VOIs, PVC and kinetic modeling.

Current results demonstrate that frame-based MC lowers between-subject variability in the striatum, which is consistent with previous reports showing that MC lowers variability [Chen et al., 2018, Montgomery et al., 2006]. In spite of these reports, many recent studies do not include MC in their preprocessing strategy and without justification (e.g., [Zientek et al., 2016, Hinderberger et al., 2016, Frick et al., 2015]). Uncorrected head motion reduces the measured activity [Jin et al., 2013], and frame-based MC without redoing the PET reconstruction may introduce attenuation correction errors. This latter correction is often neglected [van den Heuvel et al., 2003]. The relatively small effect of MC may be due to limited head motion in the data without MC, a potential consequence of subjects being carefully instructed to remain still inside the scanner, despite the long scan time. In the absence of motion, MC will lead to some degree of smoothing due to an interpolation, which may explain the reduced variability.

From a statistical perspective, each PET workflow can be considered a statistical model used to estimate the true underlying $\mathrm{BP}_{\mathrm{ND}}$. However, as all models can be characterized in terms of its bias and variance, there is a bias-variance trade-off to consider. This means, that every time variance is reduced (e.g. by performing MC) we introduce a bias in the $\mathrm{BP}_{\mathrm{ND}}$ estimate that will make it deviate from its true value. The result of our analyses in this study is an approximation of the null distribution of $\mathrm{BP}_{\mathrm{ND}}$ and between-subject variability, respectively, across subject selection, data acquisition, and preprocessing. These null distributions capture the expected value and the total variation from the applied models in the literature (i.e. PET workflows). In order to decompose the total variation into components of subject selection, data acquisition and preprocessing one can fix two of the components, while varying the third. This can be done by considering a data set of N subjects (assumed to represent the entire population) that have undergone the same PET data acquisition. This data set can then be preprocessed in various ways to not only capture the total variation of preprocessing, but also to capture the contribution from each preprocessing step. The data set can further be expanded to include a repeated measurement on the same subject, to provide the variability of repeated measures. The variability of repeated measures is important knowledge because it will reflect our ability to detect differences in binding following an intervention.

In the next Chapter, I will use a test-retest data set to evaluate the impact of a subset of the identified preprocessing choices on measures of bias, between-subject and within-subject variability. Based on the evaluation, I will provide recommendations for an optimized preprocessing strategy that is optimal for a given study design (cross-sectional or longitudinal) across all brain regions, or with an a priori hypothesis for a specific brain region. Finally, I will update Figure 4.4 with the new results of the distribution arising from the preprocessing of the group average $\mathrm{BP}_{\mathrm{ND}}$ and the corresponding standard deviation.

CHAPTER 5

Study 2: Evaluation and Optimization of Preprocessing

5.1 Introduction

This chapter is adapted from the peer-reviewed conference article Nørgaard et al., 2018c ([C]) and the submitted journal article Nørgaard et al., 2019 ([B]).

In **Chapter 4**, I showed that the impact of data acquisition and preprocessing seems to be an overlooked aspect in modern PET neuroscience with 21.150.720 available workflows in the [^{11}C]DASB-PET literature. I demonstrated that the assumption that the outcome of a PET study is insensitive to preprocessing does not hold. This suggests that there may be an advantage of identifying a preprocessing strategy that is more optimal than others across both subjects and regions. Furthermore, because different regions have differences in structure (gyrification and thickness), signal and noise, there may exist distinct preprocessing strategies that are optimal for each specific brain region.

The effects of preprocessing strategy have been investigated in numerous studies by [Montgomery et al., 2006], [Jin et al., 2013], [Schwarz et al., 2017], [Schain et al., 2014], [Greve et al., 2016], [Ichise et al., 2003] and [Ogden et al., 2007], among others, who showed that MC, co-registration, delineation of VOIs, PVC and kinetic modeling have impact on PET results. In this Chapter, I extend the work of previous studies to examine a set of commonly used preprocessing strategies from the literature, including their interactions.

The main goal of this Chapter is to implement a framework for measuring the performance of preprocessing choices, and to provide recommendations on the optimal pipeline. The performance metrics are based on reproducibility and prediction, and the framework is designed to be used in the early test-retest stage for a new radioligand, where recommendations are made to the community for subsequent studies.

The framework was used to show the primary results: **(1) there exists a pipeline that is optimal for all subjects and across brain regions**. However, **(2) there exists a heterogeneous set of region-specific pipelines that outperform the optimal pipeline for all regions**. Finally, **(3) I demonstrate that univariate and multivariate analysis models used to detect differences in $\mathbf{BP_{ND}}$ between scan sessions are preprocessing dependent.**

5.2 Methods

In this section, I establish a framework for selecting (1) an optimal pipeline suitable to all subjects and brain regions, and (2) an individually optimized pipeline for each specific region. First, I discuss the Data-Analysis Chain (Chapter 5.2.1).

5.2.1 The Data-Analysis Chain

Subject Selection: 30 healthy female participants were included in the study (mean age: 25±5.9 years, range: 18-37). Details are provided in Chapter 3.1.

Data Acquisition: All participants were PET scanned twice on separate days with the same imaging protocol. The participants received a placebo treatment between scans, and are therefore considered to represent test-retest. Details are provided in Chapter 3.2.

Preprocessing:
The preprocessing steps are listed in Chapter 3.3 in the order in which they were applied. Specific rationales for including/excluding each unique preprocessing step and their options are listed in Chapter 3.3.

Statistical Analysis:
I evaluated and optimized the preprocessing pipeline using statistical performance metrics related to reproducibility: test-retest bias, within-subject variability (WSV), between-subject variability (BSV) and the Intraclass Correlation Coefficient (ICC). The interactions of preprocessing steps in the pipeline were mea-

sured, by testing all possible combinations of MC (two choices), co-registration (four choices), delineation technique (three choices), PVC (four choices) and kinetic modeling (four choices). This resulted in $2 \times 3 \times 4^3 = 384$ combinations of preprocessing. Details are provided in Chapter 3.4.1. For the false-positive analysis, I used univariate (paired t-test) and multivariate (LDA) analysis models as described in Chapter 3.4.2.

5.2.2 Preprocessing Optimization Across Subjects and Regions

The reproducibility of BP_{ND} estimates across subjects and regions are known to be heterogeneous [Ogden et al., 2007, Zanderigo et al., 2017]. To identify an optimal preprocessing strategy across subjects and regions ("FIX" pipeline), I used a non-parametric technique with the gSNR metric to measure the performance of preprocessing, as described in detail in Chapter 3.4.1 (Figure 5.1). The technique is a conservative approach for identifying a set of optimal preprocessing strategies across subjects and regions at 95% confidence.

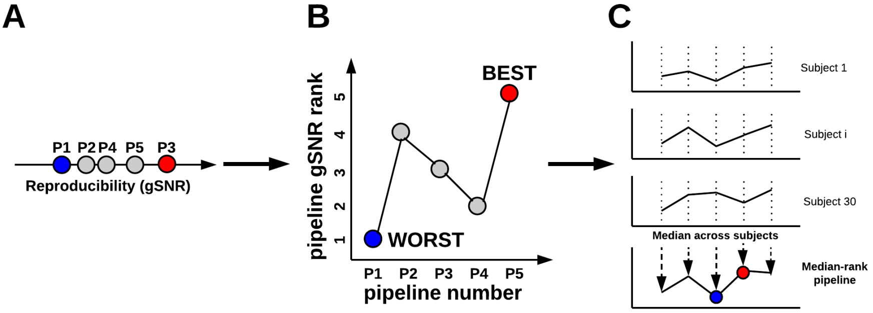

Figure 5.1: Framework to identify an optimal preprocessing pipeline across subjects and regions. (A) For subjects i=1,...,30, measure the gSNR for all pipeline combinations. (B) For each subject, rank pipelines according to the gSNR with the highest rank being the best (red) and lowest rank being the worst (blue) (C) Obtain pipeline rank profiles for all subjects, and take the median rank of each pipeline, across subjects. The significance of the median-rank profile, can be assessed using a Friedman rank test.

5.2.3 Region-Specific Preprocessing Optimization

I identified the set of preprocessing combinations that minimized the BSV and WSV, respectively, or maximized ICC for each region. To stabilize the performance metrics and to remove subject-specific artefacts I used statistical subsampling, selecting a subset of 20 subjects without replacement, and this was repeated over 1000 iterations to compute a mean estimate and a 95% confidence interval. Then, I examined the differences in performance between the optimal pipeline across subjects and regions, and the region-specific optimal pipeline. Details are provided in Chapter 3.4.1.

5.3 Results

5.3.1 Test-retest Bias

Figure 5.2 plots the percent bias between test and retest BP_{ND} as a function of preprocessing strategy in the occipital cortex. Across regions, 98% of all tested pipelines, showed a negative bias (range: -6% to 0%). This means that the BP_{ND} was lower on the second scan compared to the first scan.

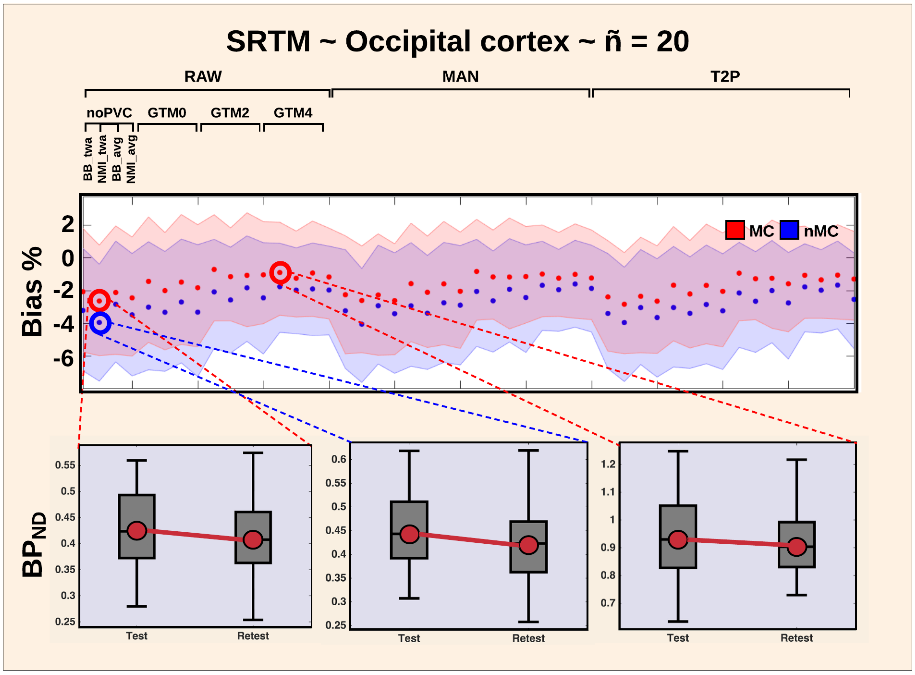

Figure 5.2: Test-retest bias (%) as a function of pipeline for the occipital cortex, when SRTM is applied. The use of motion correction generally decreases the bias (range: -1% to -4%). This is highlighted by the three plots in the bottom, showcasing the test-retest effect on BP_{ND}.

5.3.2 Within- and Between-Subject Variability

Figure 5.3A and 5.3B show the WSV and BSV across brain regions, and as a function of the preprocessing step MC (with/without). In Figure 5.3C and 5.3D the WSV and BSV are shown for noPVC vs. GTM with 4 mm. Figure 5.3E and 5.3F display the WSV and BSV for SRTM vs. MRTM2.

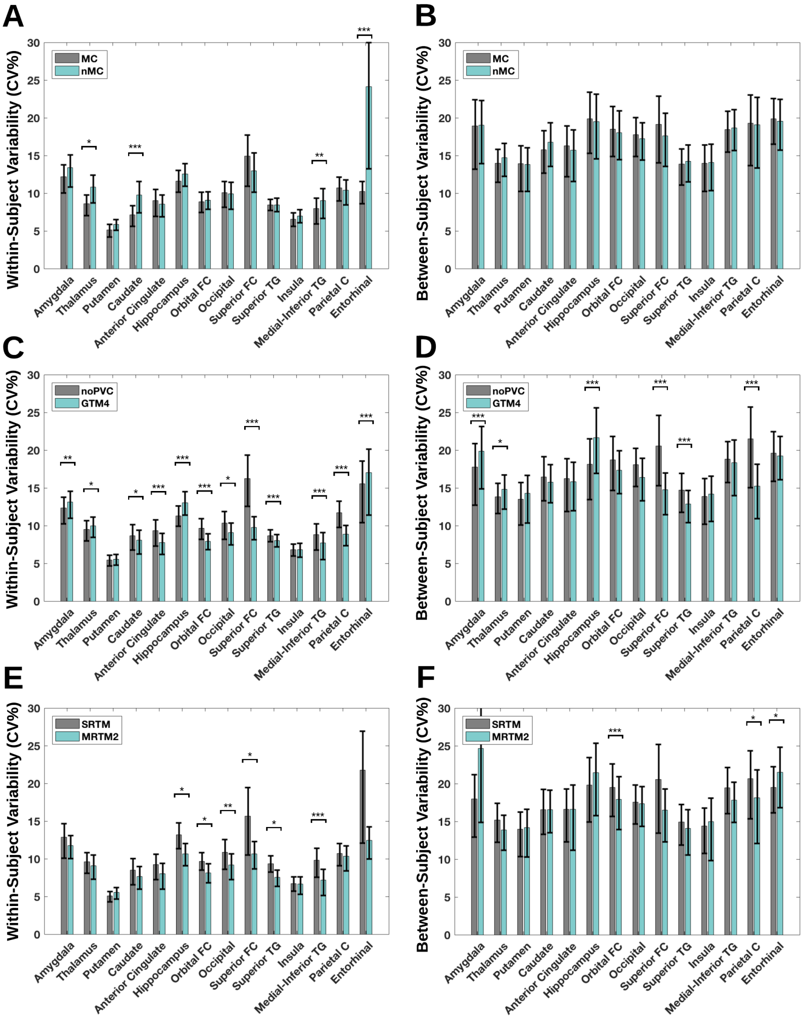

Figure 5.3: **(A-B)** within- and between-subject variability for 14 regions with or without motion correction, including a 95% confidence interval **(C-D)** similar to A and B, but with either no partial volume correction (noPVC) or with the Geometric Transfer Matrix with a 4 mm PSF (GTM4) **(E-F)** similar to A and B, but with the Simplified Reference Tissue Model (SRTM) or the Multilinear Reference Tissue Model 2 (MRTM2). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, FDR corrected for multiple comparisons (FDR=0.05).

5.3.3 Preprocessing Optimization Across Subjects and Regions

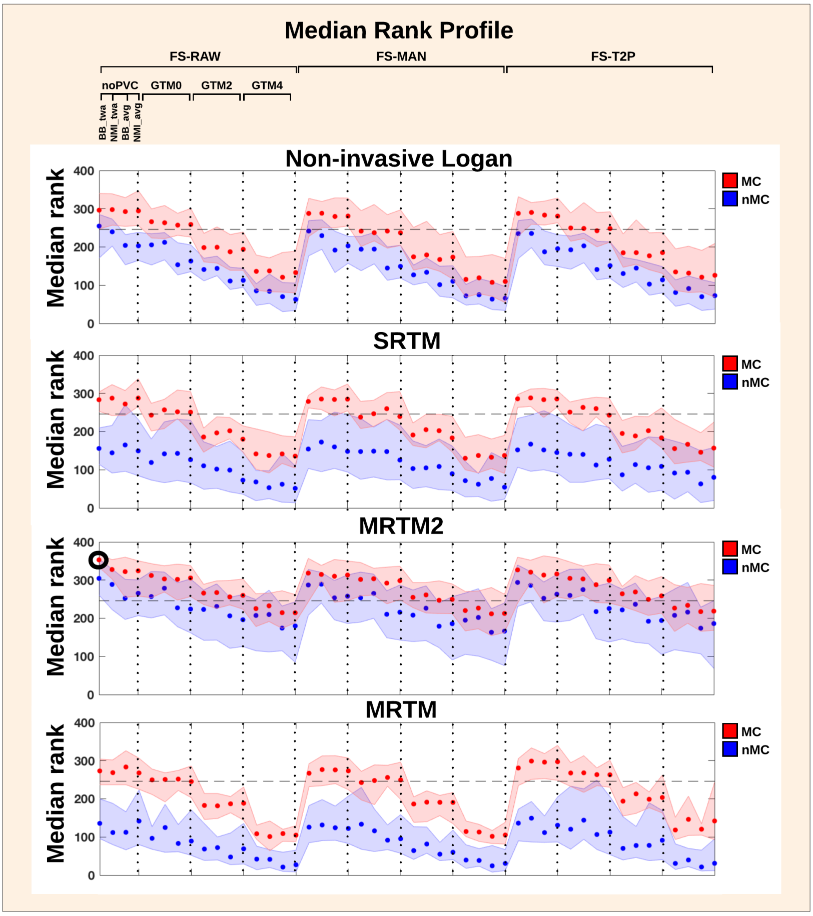

Figure 5.4: Median rank profile for all pipelines across all subjects. The shaded errorbars indicate 95% confidence intervals. The optimal pipeline across subjects and regions (FIX) is visualized by the black bold circle. The horizontal dotted line indicates that pipelines below this line are significantly different from FIX. The pipelines above the cut-off are not significantly different from each other.

The median-rank profile for the assessment of relative pipeline performance for each pipeline and across all brain regions is shown in Figure 5.4. I identified a significant pipeline effect across subjects ($p < 0.0001$, Friedman test), suggesting the existence of an optimal preprocessing pipeline across regions and subjects. The highest median rank was achieved with the preprocessing strategy: MC, BB_{TWA}, FS-RAW, noPVC, and MRTM2. I also identified a set of other pipelines, that were not significantly different than FIX (Dunn-Sidak test, corrected for multiple comparisons for all possible pairwise combinations, $p = 0.05$). The pipelines below the dotted horizontal line in Figure 5.4 are significantly different from FIX. MC consistently increased the median rank. The rank for MRTM2 with either MC or nMC were not significantly different from each other, whereas the non-invasive Logan, SRTM and MRTM showed significantly higher rank after the application of MC (non-overlapping CI's). The application of PVC generally decreased the median rank with increasing PSF. The application of PVC and MRTM2 did not affect the median rank, whereas the application of PVC with other choices of kinetic models significantly lowered the median rank. Co-registration with the time-weighted PET image marginally increased the median rank, but only when MC was not included in the pipeline. When MC was applied, the choice of co-registration only resulted in minor effects on the rank. The choice of delineation technique did not affect the rank.

5.3.4 Region-Specific Preprocessing Optimization

Table 5.1 summarizes the results of region-specific optimization, where the pipeline strategy that optimizes each performance metric for each specific region is listed. In going from FIX to an optimal region-specific preprocessing strategy, the BSV was reduced (range: 0% to 8%) in CV (mean change of 3.6±2% from FIX; p=0.0001, Wilcoxon signed rank test) with amygdala and superior frontal cortex showing the largest improvements (8% and 5%, respectively). The WSV was reduced (range: 0% to 5%) in CV (mean change of 1.7±1.56% from FIX; p=0.0006, Wilcoxon signed rank test) with ACC (3%), orbital FC (3%), superior FC (5%), and parietal cortex (4%) showing the largest reductions. Across regions, use of either MRTM or MRTM2 consistently reduced the WSV. The application of GTM with a 4 mm PSF minimized the BSV in all regions, except in the amygdala, thalamus and hippocampus. The WSV was also minimized following GTM4, except in the insula and entorhinal cortex.

Table 5.1: Overview of optimal pipelines for 14 brain regions, when optimized by median-rank (FIX), within-subject variability (WSV), between-subject variability (BSV) and intra-class correlation (ICC). 1st letter (Delineation of regions; A=FS-raw, B=FS-man, C=FS-T2p), 2nd letter (Motion Correction (MC); A=MC, B=noMC), 3rd letter (Co-registration; A=BB_{twa}), B=NMI_{twa}), C=BB_{avg}, D=NMI_{avg}), 4th letter (Partial Volume Correction (PVC); A=noPVC, B=Geometrix Transfer Matrix (GTM) 0 mm, C=GTM 2 mm, D=GTM 4 mm), 5th letter (Kinetic modeling; A=MRTM, B=MRTM2, C=SRTM, D=Non-invasive Logan).

	FIX	WSV	BSV	ICC
Amygdala	AAAAB	CBBCB	BAAAD	ABACB
Thalamus	AAAAB	BAAAA	ABBAD	BABDA
Putamen	AAAAB	CAAAA	CADDA	AABDA
Caudate	AAAAB	CAADB	CADDA	AAADB
Anterior Cingulate	AAAAB	BBADB	ABDDD	CBADB
Hippocampus	AAAAB	BBBAB	ABBAD	CBBCB
Orbital FC	AAAAB	BBBDB	CBDDD	BBADB
Occipital C	AAAAB	BABDB	ABDDC	CABDA
Superior FG	AAAAB	ABCDB	ABBDA	CBADC
Superior TG	AAAAB	BBBDB	AABDD	BBABB
Insula	AAAAB	CABBA	BABDD	CBBDB
Medial-Inferior TG	AAAAB	BBBDB	BABDB	CBBDB
Parietal C	AAAAB	ABADA	ABCDB	BBABC
Entorhinal C	AAAAB	CABAB	CBBDD	BABDB

5.3.5 False-Positive Analysis

As the data originates from a test-retest study there should be no differences in BP_{ND} between test and retest. Therefore, significant differences between test and retest are considered a false-positive. The univariate paired t-test was used to detect statistical mean differences in BP_{ND} between test and retest across regions and pipelines. The results are summarized in Figure 5.5 with/without correction for multiple comparisons using FDR, with higher FPR being worse. Without correction, 36% of the tests across regions and pipelines resulted in a significant result ($p < 0.05$). With correction, the FPR was 2.5% across regions and pipelines. The preprocessing choices that contributed to the false-positives were mainly MC in combination with the kinetic models, SRTM and MRTM.

The multivariate analysis using LDA was used for predictive classification of test (class 1) and retest (class 2) BP_{ND}. The results of the multivariate analysis are summarized in Figure 5.7. Depending on the preprocessing strategy, classification accuracies varied from 37% to 70%, with a mean accuracy of 51%. The pipeline that provided the highest classification accuracy (63.3%, $p = 0.12$) was: noMC, NMI_{AVG}, FS-T2p, noPVC, and MRTM. For this pipeline, one of the 10 repetitions of the 5-fold cross-validation resulted in a classification accuracy of 70%, and therefore significantly different from its permuted null-distribution (p = 0.01).

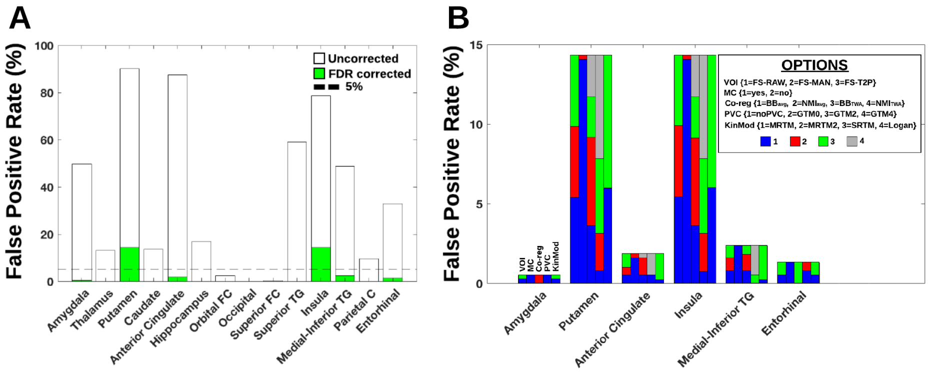

Figure 5.5: **(A)** Number of significant results (paired t-test, $p < 0.05$) in 384 pipelines divided by 384, expressed as a percentage for 14 brain regions. Blank is not corrected for multiple comparisons, whereas green is corrected using FDR. **(B)** Number of significant results (paired t-test, $p < 0.05$) 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 has 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)

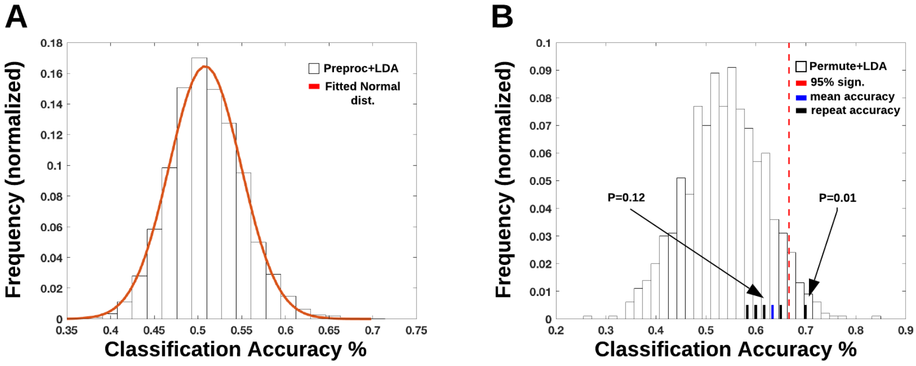

Figure 5.6: **(A)** Normalized distribution of classification accuracies (%) for 10 times repeated 5-fold cross-validation and for 384 different preprocessing choices **(B)** Normalized distribution of 1000 permuted classification accuracies (%) for the pipeline maximizing the classification accuracy 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).

5.4 Discussion

The study in this chapter establishes a comprehensive preprocessing framework for measuring the effect of preprocessing steps and their interactions on measures of bias, WSV and BSV, needed sample size and the FPR. I also established the heterogeneity of region-specific variability in response to preprocessing strategy, and present the first evaluation of interactions between MC, co-registration, delineation techniques, PVC and kinetic modeling in the quantification of binding using dynamic PET.

Independent of preprocessing strategy, I demonstrated that $\mathrm{BP}_{\mathrm{ND}}$ was lower at the second scan compared to the first. This observation has also been demonstrated previously [Kim et al., 2006], reporting a bias ranging from 2.5% to 7.5%. Two other test-retest studies using [^{11}C]DASB-PET [Frankle et al., 2004, Ogden et al., 2007] did not apply a bias metric in their examination. The bias may be the result of a true biological effect, but it may also be introduced in the data acquisition and/or preprocessing stage. If the bias is determined by biological processes it means that attempts to identify a pipeline minimizing the bias is counter productive. For example, if an intervention acts to increase the binding, the negative bias may cancel out the effect of the intervention.

Optimization of preprocessing across subjects and regions, identified a set of optimal preprocessing strategies showing significant effects for MC, co-registration, PVC and kinetic modeling. Replicating previous findings, I demonstrated that MC is an important step in the pipeline to increase reproducibility. MC has previously been shown to significantly affect PET results [Montgomery et al., 2006, Jin et al., 2013], but despite these reports, 40% of [^{11}C]DASB-PET studies left out MC in their analysis (A). While MC generally improved reproducibility across regions and subjects, I also identified a set of regions (thalamus, caudate, medial-inferior TG and entorhinal cortex) that had significantly lower within-subject variability following MC. Thalamus and caudate have often been used as high-binding regions for estimation of k_2' [Frokjaer et al., 2015, Nørgaard et al., 2017], and the variability coming from this estimation will be transferred into the kinetic models using the estimate (i.e. MRTM2 and non-invasive Logan). This will have an impact on the estimation of the $\mathrm{BP}_{\mathrm{ND}}$ in the whole brain [Ichise et al., 2003, Mandeville et al., 2016]. Putamen was the region least affected by preprocessing strategy, minimizing both WSV and BSV relative to thalamus and caudate. Therefore, to minimize potential biases originating from subject-dependent differences, the putamen as a high-binding region is suggested to be used as an optimal choice in future studies.

The performance of the rank-analysis was largely dependent on the use of noPVC or GTM, with the latter contributing negatively to the rank. These results differ from [Greve et al., 2016], who suggested that the GTM was the preferred method for VOI analysis. The cause for this discrepancy is related to a distinct difference in PVC performance across subcortical and cortical regions (Figure 5.3). The application of GTM resulted in a significant decrease in WSV in most cortical regions, whereas it significantly increased both WSV and BSV in the amygdala, thalamus and hippocampus. The effect may be attributed to PVEs being more correlated with the cerebellum in these regions, resulting in unstable estimates of $\mathrm{BP_{ND}}$ [Greve et al., 2016]. The findings highlight the utility of using multiple performance metrics over any single metric [Churchill et al., 2012, Churchill et al., 2015].

I also identified a trade-off in WSV and BSV at the group level. Minimization of BSV increased WSV relative to the FIX pipeline, particularly when applying the non-invasive Logan. Quantification using the non-invasive Logan is often preferred, as it produces a low between-subject coefficient of variation [Tyrer et al., 2016, Logan et al., 1996] at the expense of a bias. I demonstrate that the consequence of choosing non-invasive Logan to decrease BSV is a 3-5% increase in WSV (B). These findings indicate that depending on the experimental design, the choice of preprocessing should be selected with caution and with careful consideration of the study goals. The variability of the measured variable (i.e. $\mathrm{BP_{ND}}$) will also influence the statistical power of a study [Whitley and Ball, 2002]. However, while the sample size required to show an effect on a variable, is ultimately determined by the variability of the variable, studies may become underpowered if incorrect variability measures are used [Button et al., 2013].

The comparison of univariate and multivariate analysis models and their impact on the false-positive rate provided insight into the effects of preprocessing on the detection of differences between test and retest $\mathrm{BP_{ND}}$. The univariate model with varying preprocessing choices, was still able to detect significant differences, despite correcting for mulitple comparisons. Correction for multiple comparisons should therefore always be carried out to limit the inflation of false-positive results [Bennett et al., 2009]. The multivariate model with varying preprocessing choices, was not able to detect any false positives, evaluated with cross-validation and permutations. This may be caused by reduced power in the cross-validation due to splitting of the data, but it may also be due to adequate model generalizability [Varoquaux et al., 2017] compared to the overfitted univariate model. Given that the data is test-retest, there should be no detectable differences. The current results demonstrate that univariate models are more sensitive to preprocessing choice, and unless corrected for multiple comparisons, results in increased false-positive rates. Based on these results, I suggest that care must be taken in the analysis of longitudinal data to avoid attributing an effect to a treatment/-condition that was due to the retest alone.

Finally, to round of this chapter, I update Figure 4.4 from Chapter 4 with the results of the preprocessing framework obtained in the current chapter (Figure 5.7) and compare them with the literature. The results indicate that preprocessing is responsible for nearly 50% of the total variation in the average $\mathrm{BP_{ND}}$ in the striatum (Figure 5.7A) and nearly 40% of the standard deviation of the BP_{ND}. The reason for the high BP_{ND} in the striatum compared to the literature is due to the high resolution of the HRRT scanner and the subsequent application of PVC. Both methods will increase the signal.

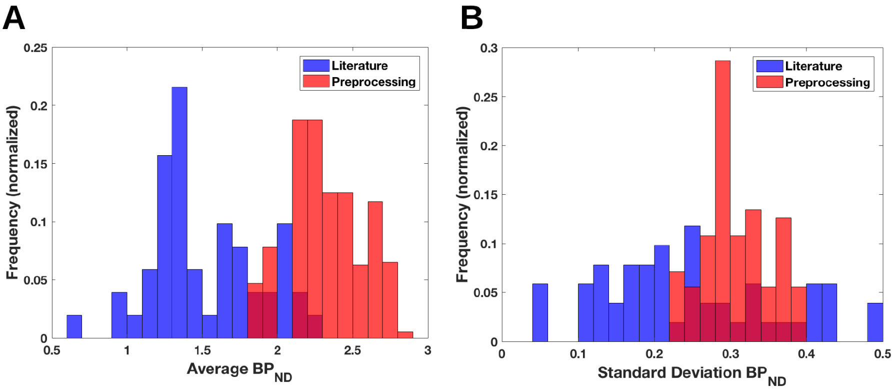

Figure 5.7: Striatal **(A)** group average BP_{ND} and **(B)** standard deviation in groups of healthy participants across 50 $[^{11}C]$DASB-PET studies (blue) and from 384 different preprocessing strategies used in the test-retest data set (red).

In this chapter, I demonstrated that **(1) there exists a set of optimal preprocessing pipelines (FIX) that adapt to all subjects and brain regions**. However, **(2) there exists a heterogeneous set of region-specific pipelines that outperform the FIX pipeline**. Finally, **(3) I show that univariate and multivariate analysis models used to detect differences in BP_{ND} between scans are preprocessing dependent, and interact to affect the false-positive rate**.

Chapter 6

Study 3: Different Preprocessing Choices Lead to Different Conclusions

6.1 Introduction

This chapter is adapted from the submitted manuscript [D]: Nørgaard M, Ganz M, Svarer C, Frokjaer VG, Greve DN, Strother SC, Knudsen GM. The Impact of Different Preprocessing Strategies in PET Neuroimaging: A [^{11}C]DASB-PET Case. Submitted to Journal of Cerebral Blood Flow and Metabolism, Jan 2019.

In the previous chapter, I demonstrated that preprocessing choices impact the test-retest bias, within-subject variability, between-subject variability, and the false-positive rate. I also demonstrated that the variation coming from preprocessing accounted for up to 50% of the total variation found in the literature and across PET centres. However, while it may be inevitable that different PET centres use different methods, the key question that remains unanswered is how these differences affect the conclusions of a study?

In this chapter, I extend the preprocessing framework presented in Chapter 5, by **(1) examining in an independent data set how the conclusions in the study depend on the choice of preprocessing strategy**. The examination of a range of reasonable preprocessing strategies should ensure that a conclusion is not driven by the result of a single pipeline.

Recently, it has been proposed that science is entering a reproducibility crisis, with limited ability to reproduce previous observations despite applying the same methodology [Baker and Penny, 2016, OpenScienceCollaboration, 2015]. In reality, a replication study is never completely overlapping with the original study, with differences in available equipment, settings and sample data [Goodman et al., 2016]. However, the generation of a plausible conclusion is often taken as justification of the methodological choices made, creating a systematic bias towards prevailing scientific expectations [Strother et al., 2002]. In this chapter, I applied the preprocessing framework established in Chapter 5 to examine which of the pipelines reproduced the main outcome from [Frokjaer et al., 2015], namely a positive association between the emergence of depressive symptoms and change in cerebral 5-HTT binding following a gonadotropin-releasing hormone agonist (GnRHa) intervention. In addition, I also tested how preprocessing strategy would influence the association between the personality trait neuroticism and change in 5-HTT binding from baseline, which was also part of the original analysis. The original study used the pipeline: with MC, $\mathrm{NMI}_{\mathrm{TWA}}$ co-registration, delineation of VOIs using PVElab, noPVC and MRTM2. Because preprocessing strategies in the [^{11}C]DASB-PET literature have been assumed to produce near similar results [Kim et al., 2006, Ginovart et al., 2001], it was hypothesized that by switching preprocessing strategy this would not affect the conclusions of the study.

6.2 Methods

6.2.1 The Data-Analysis Chain

Subject Selection: 29 healthy female participants were included in the current work (mean age: 23.3±3.3 years). Details are provided in Chapter 3.1.

Data Acquisition: Participants were PET scanned twice on separate days. The participants received a subcutaneoous injection of a GnRHa implant between scans. Details are provided in Chapter 3.2.

Preprocessing:
The preprocessing strategies used are listed in Chapter 3.3 summing to a total of 384 combinations. The VOIs used in this study were the neocortex, ACC, striatum and the midbrain. Rationale for including/excluding each unique preprocessing step and their options are listed in Chapter 3.3.

Statistical Analysis:
For each region, linear models were constructed with $\mathrm{BP}_{\mathrm{ND}}$ as the independent variable, and either neuroticism score or Hamiltons Depression score as the dependent variable. In total, this results in 3072 linear models. Linear models with p-value below 0.05 were considered statistically significant.

6.3 Results

6.3.1 Depressive Symptoms and Preprocessing

Figure 6.1 shows the distribution of p-values for the association between change in neocortical BP_{ND} from baseline and Hamilton change from baseline (Figure 6.1A). By dividing the histogram into data with MC (red) and without MC (blue), I identified a clear segregation in p-values with 36% of pipelines being significant ($p < 0.05$) and 64% non-significant. Effect sizes (i.e. Pearson's correlation) varied from 0.15 to 0.45. No significant p-values included nMC (without MC), indicating that this is a suboptimal procedure for replication. Figure 6.1B (lower) shows the association for a single significant pipeline as highlighted by the black star in Figure 6.1A. The black star is the recommended FIX pipeline from Chapter 5 (MC, BB_{TWA}, FS-raw, noPVC and MRTM2). Figure 6.1B (upper) shows how the change in BP_{ND} from baseline varies as a function of preprocessing in a single subject, demonstrating that the preprocessing variability is nearly as large as the between-subject variability.

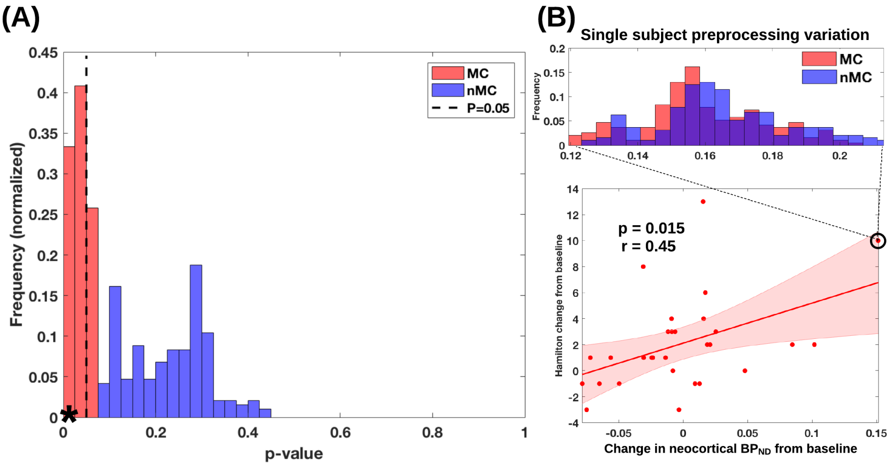

Figure 6.1: (A) Histogram of p-values obtained across 384 preprocessing strategies examining the association between change in neocortical BP_{ND} and in Hamilton score from baseline in the GnRHa group. MC = 'Motion Correction', nMC = 'no Motion Correction', **(B)** Lower plot shows the association between the change in neocortical BP_{ND} and Hamilton score from baseline ($p = 0.015$, Pearson's $r = 0.45$), using the FIX preprocessing strategy from [B] (black star in (A)). The shaded error bar indicates the 95% confidence interval. Of the 384 preprocessing strategies, 36% were significant at $p < 0.05$ and they all included MC. The black circle (B, lower) and the histogram (B, upper) illustrate the variation (between 0.12 and 0.22) in the change in neocortical BP_{ND} from baseline for a single subject, across the 384 preprocessing strategies.

6.3.2 Neuroticism and Preprocessing

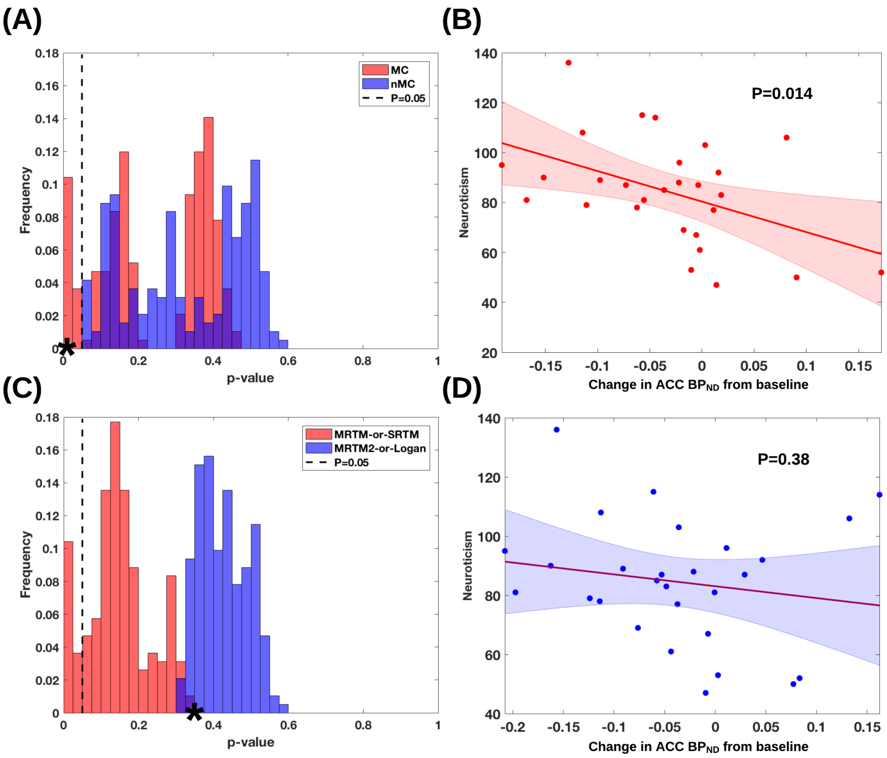

Figure 6.2: **(A)** Histogram of obtained p-values for the association between the change in ACC BP$_{\mathrm{ND}}$ from baseline and neuroticism, in the GnRHa group and across 384 preprocessing strategies. MC = 'Motion Correction', nMC = 'no Motion Correction'. **(B)** Association between the increase in ACC BP$_{\mathrm{ND}}$ from baseline and neuroticism ($p = 0.014$), using one of the 27 preprocessing strategies (black star in (A)) yielding a significant correlation ($p < 0.05$). All preprocessing strategies yielding statistically significant outcomes share the steps MC and SRTM. **(C)** similar histogram as in (A) but now divided into SRTM-or-MRTM (red) and MRTM2-or-Logan (blue) **(D)** Similar plot as in (B) but for a pipeline that generates a statistically non-significant outcome (black star in (C)).

Figure 6.2 shows the distribution of p-values for the association between the change in ACC BP$_{\mathrm{ND}}$ from baseline and neuroticism, as a function of preprocessing strategy. I identified a distinct segregation in the distribution of p-values (Figure 6.2C) between the use of MRTM-or-SRTM (red) and MRTM2-or-Logan (blue). The significant results indicate a negative association between neuroticism score and change in ACC BP$_{\mathrm{ND}}$ from baseline.

6.4 Discussion

This chapter presents the first comprehensive analysis to examine the effects of several preprocessing interactions on the outcome of a dynamic PET study. While MC was found to be the main component for replicating the original study, the results of this chapter also demonstrate interactions between other steps of the preprocessing pipeline. In particular, SRTM or MRTM combined with MC resulted in the identification of a negative association between change in ACC $\mathrm{BP}_{\mathrm{ND}}$ from baseline and neuroticism, a result that was not reported in the original study. It is therefore important to consider and to declare preprocessing strategies before analyzing the data, as different preprocessing strategies may lead to different conclusions.

In previous chapters, I demonstrated that MC is a key step in the preprocessing pipeline, replicating previous findings [Montgomery et al., 2006, Jin et al., 2013]. I also demonstrated that except from MC, PVC and kinetic modeling were the most prominent components affecting both the within- and between subject variability. In the current study, I replicate that MC (Figure 6.1) and kinetic modeling (Figure 6.2) have important effects on the results. Notably, the combination of SRTM/MRTM and MC resulted in a significant association between neuroticism and 5-HTT binding in the ACC (Figure 6.2). Only two previous studies have used the combination of nMC and SRTM [Nogami et al., 2013, Ogawa et al., 2014], whereas the remaining studies in the literature used MC before applying SRTM [Comley et al., 2013, Turkheimer et al., 2012, Abanades et al., 2011, Hammoud et al., 2010]. SRTM estimates the $\mathrm{BP}_{\mathrm{ND}}$ using non-linear least squares optimization, and it is likely that artefacts from subject-specific head motion may result in an unstable solution. Another notable observation was that the preprocessing variability as visualized in the histogram in Figure 6.1B (upper) was nearly as large as the between-subject variability (Figure 6.1B, lower). This questions whether the identified association is the result of a true subject-specific effect or if it is due to the preprocessing.

I also identified a subset of preprocessing pipelines that produced a significant negative association between neuroticism and change in ACC $\mathrm{BP}_{\mathrm{ND}}$ from baseline. Neuroticism has consistently been implicated in depression and 5-HTT levels [Tuominen et al., 2017, Hirvonen et al., 2015]. However, there may also be some characteristics of neuroticism as a trait that could potentially affect the PET measurements when scanned twice. These include that subjects with low neuroticism are likely to have increased stress levels at their first PET scan, whereas stress levels may drop on their second scan due to familiarity with the environment. The circulation of cortisol has been shown to be elevated at high stress levels, and this has been shown to increase 5-HTT synthesis [Kim et al., 2006]. Stress level may therefore affect cerebral 5-HTT levels. Consistent with this explanation, I demonstrated in Chapter 5 that cerebral 5-HTT levels were lower when healthy volunteers were scanned the second time. Since the GnRH intervention would increase vulnerability to high stress levels, one could speculate that the intervention would be associated with similar or even higher stress levels at the second scan. I tested this hypothesis, by carrying out a post-hoc exploratory analysis seeking for a potential group interaction effect between neuroticism and $\mathrm{BP}_{\mathrm{ND}}$. The interaction was identified (Figure 6.3) for some but not all regions and preprocessing choices. The results suggest that the test-retest bias may be both state- (stress level) and trait-dependent (neuroticism) [Cannon et al., 2006, Hornboll et al., 2018]. I also considered if higher stress levels would be associated with higher levels of head motion, but I did not identify any differences in motion between scans (data not shown).

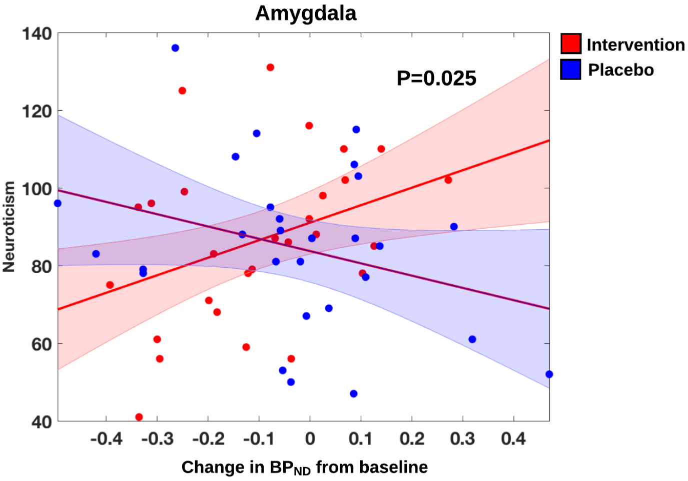

Figure 6.3: Group interaction effect between neuroticism and change in $\mathrm{BP}_{\mathrm{ND}}$ in the amygdala. Blue is the placebo group and red is the intervention group. Shaded error bars are 95% confidences intervals.

While current results give rise to interesting explanations, it is equally important to test and validate the results in an independent data set. Furthermore, there are some statistical considerations that could help researchers mitigate towards a more predictive and replicable science. This would have been obtained in the current study by the application of a predictive model evaluated with cross-validation (random effect model) instead of a linear regression model (fixed effect model). As demonstrated in Chapter 3 and 5, predictive models provide a predictive accuracy of a models ability to correctly predict the experimental condition in an independent data set [Varoquaux et al., 2017]. Nevertheless, a plausible explanation (using a fixed effect model) is often chosen over predictive accuracy, limiting the generalizability to an independent sample [Yarkoni and Westfall, 2017].

To increase generalizability, the current framework may also be used to estimate the expected conclusion of a study conditioned over multiple preprocessing strategies. The expectation provides a confidence in the extent to which the outcome of a study is valid under a set of plausible preprocessing strategies. This should help to control the probability that the conclusion could arise under the null (false-positive rate), and to verify that an outcome is not the result of a single pipeline. However, there are also some statistical limitations that need to be mentioned. For example, (1) the expectation is not sufficiently correcting for the number of tested preprocessing strategies, nor (2) does it examine whether preprocessing strategies are significantly different from each other. Finally, (3) the subset of 384 preprocessing strategies of all possible strategies, does not allow us to infer whether the expectation conditioned over preprocessing strategies is biased. As shown in Chapter 4, there exists at least 21.150.720 potential PET workflow possibilities, so it is likely that the estimated sampling distribution does not represent the true underlying distribution. Based on the observations of generalizability and prediction, it would be of great value to the community to develop a statistical framework that (1) includes a predictive component, (2) can correct for the number of tested pipelines, and (3) has no distributional assumptions (i.e. non-parametric).

In this chapter, I demonstrated that **(1) different preprocessing strategies lead to different conclusions, that further support the results in Chapter 5**. However, **(2) the limitations of a plausible explanation may be overcome by developing a predictive framework that can provide a predictive accuracy of the generalizability of the results.**

Chapter 7

Study 4: Predictive Framework to Correct for Multiple Preprocessing Options

---

7.1 Introduction

This chapter is adapted from the submitted manuscript [E]: Nørgaard M, Ozenne B, Svarer C, Frokjaer VG, Ganz M. Preprocessing, Prediction and Significance. Framework and Application to Brain Imaging. Submitted to Medical Image Computing and Computer Assisted Intervention (MICCAI), Jan 2019.

In the previous chapter, I demonstrated that different preprocessing strategies lead to different conclusions. Furthermore, I argued that the drawbacks of a plausible neurobiological explanation using a correlational analysis may be overcome by adopting a predictive point of view to increase generalizability [Gabrieli et al., 2015, Yarkoni and Westfall, 2017]. This was motivated by limitations in earlier work (Chapters 5 and 6), and a rising concern in the scientific community on the validity and reproducibility of scientific studies [Ioannidis, 2005, Simmons et al., 2011, Gelman and Geurts, 2014], and especially in neuroimaging [Button et al., 2013, Poldrack et al., 2017]. Recent efforts on data sharing (e.g. OpenNeuro.org) are now enabling researchers to open up the subject selection- and data acquisition components of the data analysis chain, by sharing raw image data publicly. In combination with major neuroimaging software packages (e.g. SPM, FSL, AFNI and FreeSurfer) providing statistical analysis tools, the outputs of statistical analyses can also be shared (e.g. NeuroVault.org). However, while the statistical methods have been under intense scrutiny in the last decades [Eklund et al., 2016, Carp, 2012a], the influence of preprocessing on the outcome of the data analysis has besides a few initiatives in fMRI [Carp, 2012a, Churchill et al., 2015] been a largely overlooked factor. Many neuroimaging centers have set up standardized preprocessing pipelines that are used for all their studies, and large multi-center collaborations such as the Human Brain Project (HBP) have implemented their own pipeline[1] that is used daily to extract features from neuroimaging studies. Pre-registration of complete data analysis workflows before carrying out a study has been proposed as a potential solution (e.g. AsPredicted.org) for limiting the "researcher degrees of freedom" [Poldrack et al., 2017]. The argument for this reasoning is that researchers should not be constrained to a single preprocessing strategy, but should at least pre-register their choice before carrying out a study. Furthermore, there might not even exist a single strategy that is optimal across studies and neuroimaging centres. Indeed, there is evidence that different workflows are optimal for different studies and individuals [Churchill et al., 2015]. However, it also seems unlikely that out of thousands of workflows, only the pre-registered one would be able to show the true biological effect. Instead, it seems more likely that a range of plausible preprocessing strategies would result in the same conclusion, as was also highlighted in Chapter 6. In the case of a strong effect, one might even hope, that irrespective of preprocessing strategy, most preprocessing strategies in combination with a statistical analysis would be able to detect the effect. Hence, it is of interest not only to identify the variance in the preprocessing [Carp, 2012a, Churchill et al., 2015], but to take one step further and identify the variance that different preprocessing strategies add to the statistical analysis and the following conclusions. By combining the different stages (i.e. subject selection, data acquisition, preprocessing and statistical analysis) into a single framework, we can identify spurious findings due to a specific preprocessing strategy, as most strategies would not be able to produce the same result. Furthermore, it also provides strong evidence for an effect, if all preprocessing strategies arrive at the same conclusion.

In this chapter, I present a non-parametric predictive framework to examine the influence of multiple preprocessing strategies on the subsequent statistical analysis. I demonstrate how the choice of preprocessing can affect our belief in the available sample data, $\boldsymbol{x}$, with class labels y, to generalize to the true underlying joint distribution $p(\boldsymbol{x}, y)$. My approach adopts a range of plausible preprocessing choices as a generative model for $\boldsymbol{x}$, and estimates the predictive performance for the conditional distribution $p(y|\boldsymbol{x})$ using permutations [Nichols and Holmes, 2002]. By permuting across preprocessing choices, the framework provides a probability 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 [Golland and Fischl, 2003]. First, I detail the framework and then (2) give an example of its application based on the same data and preprocessing as in Chapters 5 and 6 [Frokjaer et al., 2015].

---

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

7.2 Methods

The framework that I am proposing can roughly be broken into four major components, **(A)** definition of a subset of reasonable 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 **(C)** estimation of the statistical significance of the predictive accuracy (Figure 7.1).

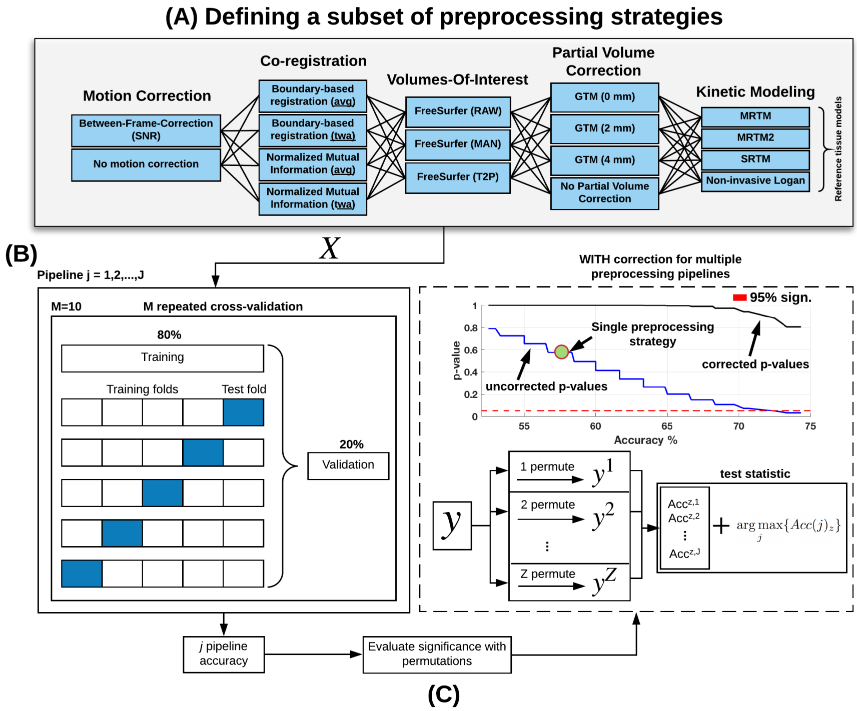

Figure 7.1: **(A)** Definition of a subset of preprocessing strategies $j = 1, ..., J$: This includes preprocessing steps such as motion correction, co-registration, delineation of volumes of interest, partial volume correction, and kinetic modeling. **(B)** Model selection and cross-validation: For each pipeline j, select a classification model (e.g. Linear Discriminant), and a nested cross-validation scheme with M repetitions, 80% training data, and 20% validation data. **(C)** Evaluate 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 preprocessing pipelines and for Z permutations, generate a null-distribution of maximal accuracies across preprocessing pipelines. 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 randomly permuted null-distribution at a significance level α.

7.2.1 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, where $\boldsymbol{x_n} \in \mathbb{R}^p$ are observations with p features and $y_n \in \{-1, 1\}$ are the corresponding class labels, the data is commonly preprocessed using a set of steps such as motion correction, co-registration and partial volume correction. Designing the most optimal sequence of 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 $\{(\mathbf{x}_{n,j}, y_n)\}_{n=1}^{N}$.

7.2.2 Model Selection and Cross-validation

Once the data has been preprocessed it is ready for statistical analysis. Next, we need to (1) select a model and tune the model parameters to the data, and (2) assess the chosen predictive model by estimating the future prediction ability of the model. For both (1) and (2), one common approach is to use cross-validation and evaluate the model in an independent test set. 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) using e.g. 5-fold cross-validation. The validation data has to be independent of the training data and completely held out of the training procedure. Additionally, the procedure has to be repeated so that each observation is assigned to the validation data exactly once. 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 [Varoquaux et al., 2017].

7.2.3 Permutation Test for a Single Strategy

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. The significance of each model and pipeline is estimated by randomly permuting the class labels Z times (i.e. sampling a permutation π^z from a uniform distribution over the set, $\mathbf{\Pi}_N$, of all permutations of indices $1, ..., N$) and re-running the above M times repeated 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 [Varoquaux et al., 2017], by introducing a test statistic of maximal accuracies across preprocessing pipelines. This approach should have a strong control over experiment-wise type I error [Nichols and Holmes, 2002].

7.2.4 Permutation Test for Multiple Strategies

Rather than computing the permutation distribution of the accuracy for a given preprocessing pipeline j, we compute the permutation distribution of the maximal 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,
  - save the maximal accuracy across pipelines J,

    $$t^z_{max} = \arg\max_j \{Acc(\mathbf{x}_{1,j}, y_{\pi^z_1}, ..., \mathbf{x}_{N,j}, y_{\pi^z_N})\}$$
- Construct an empirical cumulative distribution of max 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 actual labels for each pipeline j, $t_{0,j} = Acc(\mathbf{x}_{1,j}, y_{1,j}, ..., \mathbf{x}_{N,j}, y_N)$, and its corresponding p-value p^j_0 under the empirical distribution $\hat{P}_{max}$.

The null hypothesis assumes that the two classes have identical distributions,

$$\forall \boldsymbol{x} : p(\boldsymbol{x}|y = 1) = p(\boldsymbol{x}|y = -1).$$

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 maximal accuracy. We can reject the null hypothesis for any preprocessing pipeline with an accuracy exceeding this threshold.

7.3 Experiments

I illustrate the use of the framework in a single experiment, using 31 healthy female participants [Frokjaer et al., 2015]. Details are described in Chapter 3.1. All participants were PET scanned twice on separate days, and all participants received a GnRH intervention between scans. Details are provided in Chapter 3.2. The data, $\mathbf{x}$, consists of 60 observations (29 paired observations, 1 baseline and 1 intervention scan) each with levels of 5-HTT BP_{ND} in 34 cortical brain regions covering the entire neocortex for each preprocessing pipeline. For quantification of BP_{ND}, I used 384 combinations of preprocessing. Details are provided in Chapter 3.3. For statistical analysis, I used an LDA model to train a classifier to predict the classes (baseline and intervention), and jackknifing (sampling without replacement) for cross-validation. Details are provided in Chapter 3.4 The cross-validation was iterated 10 times to obtain an unbiased mean estimate of the accuracy, and the number of permutation iterations was 1,000. To obtain true independence between the sample data and the class labels in the cross-validation, samples for each subject (both scans) were always paired.
In Figure 7.2, I show the obtained predictive accuracy and corresponding p-value, as a function of preprocessing strategy. Every point on the blue line and the black line is a preprocessing pipeline (Figure 7.2). The predictive accuracy varies from 52% to 75%, when switching preprocessing strategy. There also exists a set of preprocessing strategies that are significantly different ($p < 0.05$) from their permuted null distribution (blue line in Figure 7.2). The black line in Figure 7.2 shows the p-values corrected according to the maximal permuted null distribution. After correction, a much higher accuracy is needed in order to obtain statistical significance.

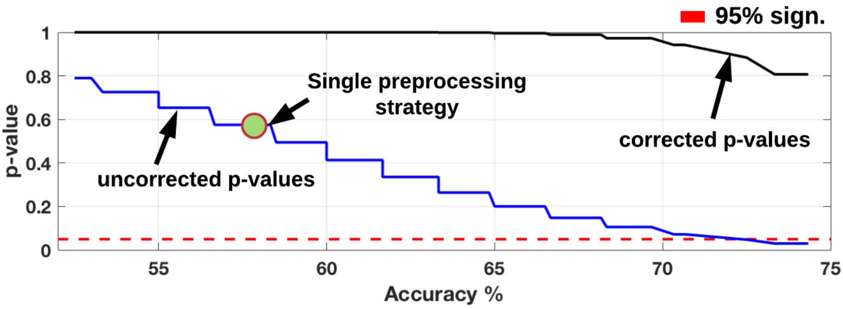

Figure 7.2: Accuracy (%) as a function of p-value for 384 preprocessing strategies. The blue line indicates the p-values according to their permuted null distribution (uncorrected) and the black line indicates the p-values according to the maximal permuted null distribution (corrected). The red dotted line is the 95% significance level.

Figure 7.3 shows the frequency of accuracies for the mean accuracies using the true labels (red distribution), for the randomly permuted labels (green distribution), and the maximal accuracies for the randomly permuted labels (blue distribution). The majority of preprocessing strategies produce accuracies that fall within the permuted null distribution, but a set of preprocessing strategies have p-values less 0.05 (i.e. less than 5% chance of observing better than 75% accuracy, given that true independence between data and labels exists). To reject the null hypothesis under the maximal accuracy, an expected accuracy of 85% is needed to obtain statistical significance at $\alpha = 0.05$ (Figure 7.3).

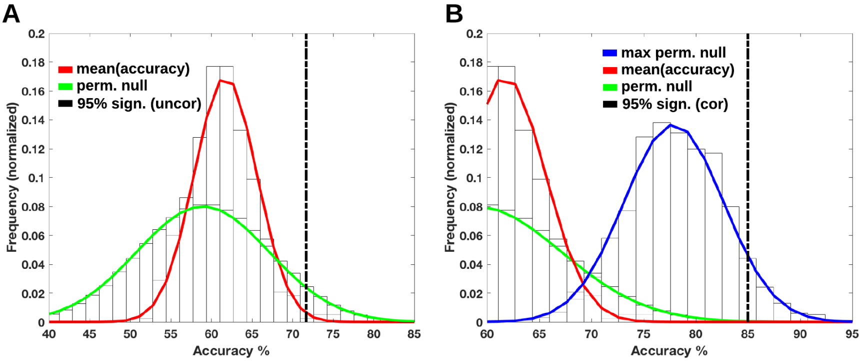

Figure 7.3: **(A)** Average classification accuracies across preprocessing pipelines obtained using nested cross-validation with 10 repeats (red). The permuted null distribution of classification accuracies (1000 permutations) across preprocessing pipelines is visualized by the green distribution. The 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 maximal classification accuracies across preprocessing pipelines. The vertical dotted line is the 95% significance level for the permuted null distribution of maximal accuracies.

7.4 Discussion

In this work, I extend the non-parametric testing of statistical significance in predictive modeling by including a plausible set of preprocessing strategies, and demonstrate its application in a pharmacological intervention study using PET. I show that for a subset of preprocessing strategies, significant predictions are obtainable, but with the majority of strategies resulting in a non-significant prediction. When correcting the p-values according to the permuted distribution of maximal accuracies, no predictions are rejected from the null hypothesis. While the predictive analysis for each preprocessing strategy is carried out according to community guidelines [Varoquaux et al., 2017, Gabrieli et al., 2015], a minority of strategies can still result in a significant prediction by chance. This may be due to the preprocessing introducing spurious relations between the sample data and the class labels, overestimating the generalizability of the predictive model. My approach also enables the examination of the predictive accuracy across multiple preprocessing strategies, providing the variance of the prediction across strategies. Based on this examination, it is advised that care must be taken when attributing an effect to a treatment/condition that was due to a single preprocessing strategy and/or predictive model. The framework that I am proposing is very flexible, and may be expanded to include more preprocessing strategies, more features (e.g. behavioural or genetic data), or other predictive models with varying model complexities. However, because the permutation part eliminates all signal, the inclusion of more preprocessing strategies will broaden the permuted null distribution due to increased noise. Therefore, an increase in the number of pipelines and by including "bad" pipelines, will punish the ability to obtain statistical significance for any pipeline. The proposed framework, provides the variance of the results across multiple preprocessing strategies, and may be used to overcome some of the limitations associated with pre-registration [Poldrack et al., 2017]. Because data acquisition is the most costly part of any experiment (i.e. the cost of a PET scan is 3000 USD), spending resources on computing power by employing the proposed framework, is negligible in comparison.

In this chapter, I demonstrated that **(1) different preprocessing strategies affect the generalizability of the results**, **(2) while nested cross-validation is considered the gold-standard to avoid circularity bias in predictive models, I show that this may be overcome by switching preprocessing strategy to obtain statistical significance**, and **(3) the resulting p-values can be corrected according to the number of tested preprocessing strategies by introducing a test statistic of maximal accuracies across strategies.**

CHAPTER 8

Conclusions and Future Work

8.1 Summary

This thesis examined the importance of improving signal detection in dynamic PET imaging studies by evaluating and optimizing choices in the preprocessing pipeline. A major finding was that preprocessing choices interacted with other stages of the data-analysis chain to affect the results (subject selection and statistical analysis).

In **Chapter 4**, I reviewed the $[^{11}C]$DASB-PET literature to highlight the variety of ways researchers have conducted their studies, while implicitly expecting generalizable results. The review provides evidence that the foundation for selecting a given preprocessing strategy seems to be an overlooked aspect in modern PET neuroscience. Based on the results, I suggest that a thorough testing of pipeline performance is necessary to increase reproducibility and to avoid biased results. The results are published in Nørgaard et al., 2018a [A].

In **Chapter 5**, I evaluated the effects of preprocessing optimization for motion correction, co-registration, delineation technique, partial volume correction and kinetic modeling. Across subjects and brain regions, a set of optimal preprocessing choices significantly improved the reproducibility. However, for each region there was a heterogeneous set of preprocessing choices that outperformed these optimal pipelines to reduce within- and between-subject variance. The false-positive rate was also shown to be dependent on both preprocessing and statistical analysis model. The results are published in Nørgaard et al., 2018b [B] and Nørgaard et al., 2019 [C].

In **Chapter 6**, I expanded the framework to investigate in an independent sample how the choice of preprocessing affected the conclusions of a study. I demonstrated that MC and kinetic modeling were critical for the conclusions drawn, and that only 36% of the tested pipelines replicated the originally reported finding. The results are submitted for publication [D].

In **Chapter 7**, I developed a non-parametric predictive framework for estimation of statistical significance, correcting for the number of tested preprocessing pipelines. This was motivated by limitations in earlier work on preprocessing evaluation and optimization in Chapters 5 and 6. The framework adopts permutation tests and cross-validation, to estimate how likely we are to obtain a significant classification accuracy, tested over all possible pipelines. While nested cross-validation should avoid circularity bias, I demonstrate that statistical significance can be obtained by switching preprocessing strategy. Without correction, the outcome will be an interaction between preprocessing and the classifier, identifying a pattern that randomly happened to correlated with the class labels. The statistical framework is submitted for publication [E].

8.2 Future Work

This section discusses future extensions of the presented research, including some of the preliminary results. The proposed future research includes: (1) surface-based and voxelwise optimization of preprocessing strategies, (2) analysis of false-positive rates in voxel-wise and surface-based PET data, and (3) data sharing.

8.2.1 Voxelwise and Surface-Based Preprocessing

This section is adapted from the manuscript in preparation : ***Nørgaard M****, Ganz M, Svarer C, Frokjaer VG, Strother SC, Knudsen GM, Greve DN. Optimizing Voxelwise and Surface-based Preprocessing Pipelines for PET Data. In preparation.*

The variability in regional PET results arising from preprocessing poses a major challenge in neuroimaging to identify areas that show an effect. When no strong a priori hypothesis of the location of the effect exists, it is common to use an exploratory voxelwise or surface-based analysis. However, noise at the single voxel is much higher than the noise in a region, making careful attempts to optimize preprocessing a requirement for successful application (Figure 8.1). The bias and variance trade-offs as a function of preprocessing are thus largely unknown for these types of analyses, and only a few papers have attempted to address some of these challenges for PET (e.g. [Greve et al., 2014]) and fMRI (e.g. [Churchill et al., 2015, Eklund et al., 2016]). In this work, I extend the framework in Chapter 5 to include a voxelwise and surface-based component, and evaluate the data using the same performance metrics. Except from noise management, there are also some technical challenges that are important to mention. For example, the memory size of a single-subject PET data set is approximately 1GB. If the data needs to undergo 384 different preprocessing strategies (or more), the size of the data set will increase to 384GB. This means that for a 60 subject data with two scans each, will result in 46TB of needed memory. Furthermore, from a computing point of view, the high dimensionality of the voxelwise (= 100.000 voxels) and surface-based (= 145.000 vertices) data, necessitates the use of high-performance- and parallel computing to speed up the analyses. Otherwise, the researcher would spent months if not years trying to search for an optimal pipeline, which is not a realistic scenario.

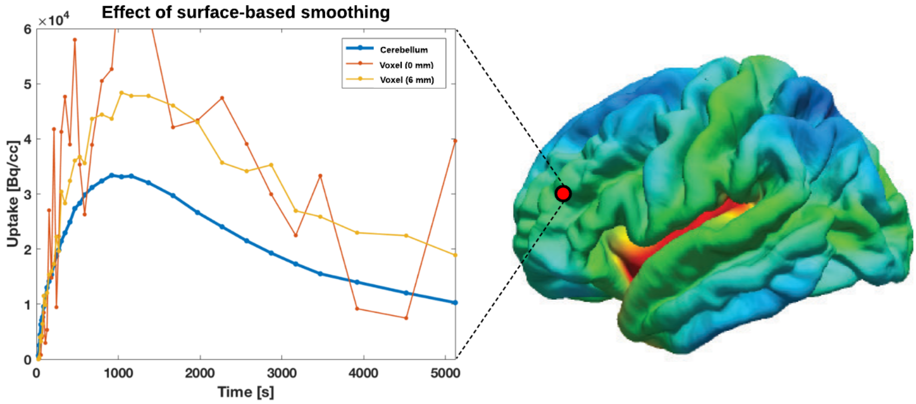

Figure 8.1: Time Activity Curve in the cerebellum (blue), for a single vertex without smoothing (orange), and for the same vertex (yellow) using surface-based smoothing with a 6 mm filter.

8.2.2 False-Positive Rates in PET

This section is adapted from the manuscript in preparation : Ganz M, ***Nørgaard M****, Beliveau V, Knudsen GM, Greve DN. False Positive Rates in Positron Emission Tomography. Presented at the Neuro Receptor Mapping meeting in London 2018. Manuscript in preparation.*

In this work we seek to investigate the false-positive rate (FPR) of whole brain PET data. This topic has recently received significant attention in the fMRI [Eklund et al., 2016] as well as the structural MRI community [Greve and Fischl, 2017], however the effects in voxelwise and surface-based PET data are largely unknown. In this work, we used PET data from 188 healthy controls imaging either the 5-HTT ([^{11}C]DASB; N = 102) or the 5-HT4 receptor ([^{11}C]SB207145; N = 86), ([Beliveau et al., 2017]). We evaluated the FPR in the PET data under random group assignments that should yield no significant results using common corrections for multiple comparisons in voxelwise as well as surface-based analyses (Figure 8.2).

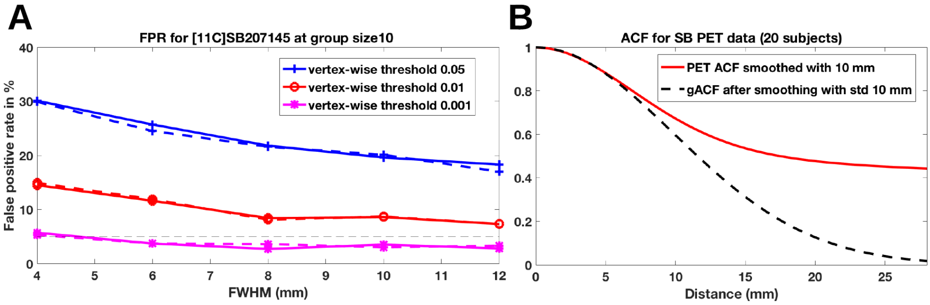

Figure 8.2: **(A)** Clusterwise FPR (%) versus applied smoothing level for Monte Carlo simulation (MCZ) for the tracer [^{11}C]SB207145 on the left (dashed) and right (solid) hemisphere for different vertex-wise statistical thresholds. The black dashed line indicates the ideal 5% FPR value **(B)** Residual autocorrelation function (ACF) for left hemisphere for 20 subjects from the [^{11}C]SB207145 data set.

8.2.3 Data Sharing

In recent years, the importance of data sharing has increasingly been recognized by the neuroimaging community, ranging from MRI, fMRI, EEG and PET. This movement comes as an acknowledgement of the substantial investment needed to acquire neuroimaging data as well as an increasing concern about the quality and heterogeneity of data across sites. In addition, both national and international funding agencies such as The Danish Research Council and National Institutes of Health are now demanding that data from the research projects they fund are stored and potentially made available for other scientists. A few important initiatives, e.g., the Human Connectome Project, have spearheaded data sharing in the MRI community, with acquisition and data analysis standards now being openly available (e.g. COBIDAS) and data sharing platforms being created (e.g. OpenfMRI) together with standardized Brain Imaging Data Structures (http://bids.neuroimaging.io). Compared to other imaging modalities the acquisition of PET data is costly, therefore the sample size of individual PET studies is in most cases relatively small (10-30 subjects). Unfortunately, with a limited sample size these statistical findings can be questionable [Button et al., 2013]. Data sharing is a cost-effective way to enlarge the sample size, and additionally the best way to make full use of research funding. From a [^{11}C]DASB-PET perspective, the majority of scans has been healthy controls, and it seems quite unethical that we do all these healthy control scans when there are hundreds of scans available. Therefore, I truly believe that data sharing and the standards required to support it are essential for expedited translation of research results into knowledge and procedures to improve human health.

8.3 Conclusions

Dynamic PET is a novel tool to non-invasively image the distribution of biochemical and pharmacological processes in the living human brain. For many years, PET centres or even individual scientists have applied and optimized their own unique preprocessing strategy, and this has been done with limited explicit knowledge of the exact impact of their choices. Although, new methodological improvements are continually being developed and refined, there have only been few comparisons between techniques using quantitative frameworks and benchmark data sets.

The current work establishes a comprehensive framework for examining the impact of a wide range of preprocessing choices in dynamic PET data. Using this framework, the variability in young and healthy subjects was found to be very sensitive to preprocessing, emphasizing the relative importance of selecting preprocessing with great caution. It is likely that this importance will be even more pronounced in subjects that are older and/or ill, due to cortical thinning and other structural/functional abnormalities. The framework can relatively easy be expanded to include more preprocessing strategies and different data, but requires an advanced data- and computing infrastructure, as computational time and needed memory will increase substantially. In addition, data sharing is needed to further expand the analysis on variability by including inter-site variability, and differences in subject selection and data acquisition. My future plan is to make the framework publicly available so all researchers can use it.

The ability to optimize preprocessing also has practical implications for future studies. The optimization framework can be used to measure signal that is obscured by suboptimal processing. This is particularly important when combining data from different PET centres, or when deciding what sample size is necessary in a given study. I strongly believe that the current work, even in the presence of relatively small sample sizes, will lead to more reproducible research outcomes.

Bibliography

[Abanades et al., 2011] Abanades, S., van der Aart, J., Barletta, J. a. R., Marzano, C., Searle, G. E., Salinas, C. a., Ahmad, J. J., Reiley, R. R., Pampols-Maso, S., Zamuner, S., Cunningham, V. J., Rabiner, E. a., Laruelle, M. a., and Gunn, R. N. (2011). Prediction of repeat-dose occupancy from single-dose data: characterisation of the relationship between plasma pharmacokinetics and brain target occupancy. *Journal of Cerebral Blood Flow and Metabolism*, 31(3):944–952.

[Aksoy et al., 2011] Aksoy, M., Forman, C., Straka, M., Skare, S., Holdsworth, S., Hornegger, J., and Bammer, R. (2011). Real-time optical motion correction for diffusion tensor imaging. *Magnetic Resonance in Medicine*, 66(2):366–378.

[Andrews-Shigaki et al., 2011] Andrews-Shigaki, B. C., Armstrong, B. S. R., Zaitsev, M., and Ernst, T. (2011). Prospective Motion Correction for Magnetic Resonance Spectroscopy Using Single Camera Retro-Grate Reflector Optical Tracking. *Journal of Magnetic Resonance Imaging*, 33(2):498–504.

[Anton-Rodriguez et al., 2010] Anton-Rodriguez, J. M., Sibomana, M., Walker, M. D., Huisman, M. C., Matthews, J. C., Feldmann, M., Keller, S. H., and Asselin, M. (2010). Investigation of motion induced errors in scatter correction for the HRRT brain scanner. In *IEEE Nuclear Science Symposuim Medical Imaging Conference*, pages 2935–2940.

[Azmitia, 1999] Azmitia, E. C. (1999). Serotonin neurons, neuroplasticity, and homeostasis of neural tissue. *Neuropsychopharmacology*, 21(2 Suppl):33S–45S.

[Bailey et al., 2005] Bailey, D. L., Townsend, D. W., Valk, P. E., and Maisey, M. N. (2005). *Positron Emission Tomography: Basic Sciences.* Springer.

[Baker and Penny, 2016] Baker, M. and Penny, D. (2016). Is there a reproducibility crisis? *Nature*, 533(7604):452–454.

[Baldassarre et al., 2017] Baldassarre, L., Pontil, M., and Mourao-Miranda, J. (2017). Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding. *Frontiers in Neuroscience*, 11:62.

[Belanger et al., 2004] Belanger, M.-J., Mann, J., and Parsey, R. V. (2004). OS-EM and FBP reconstructions at low count rates: effect on 3D PET studies of [11C]WAY-100635. *NeuroImage*, 21(1):244–250.

[Beliveau et al., 2017] Beliveau, V., Ganz, M., Feng, L., Ozenne, B., Hojgaard, L., Fisher, P. M., Svarer, C., Greve, D. N., and Knudsen, G. M. (2017). A high-resolution in vivo atlas of the human brain's serotonin system. *Journal of Neuroscience*, 37(1):120–128.

[Beliveau et al., 2015] Beliveau, V., Svarer, C., Frokjaer, V. G., Knudsen, G. M., Greve, D. N., and Fisher, P. M. (2015). Functional connectivity of the dorsal and median raphe nuclei at rest. *NeuroImage*, 116:187–195.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bennett et al., 2009] Bennett, C. M., Wolford, G. L., and Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience*, 4(4):417–422.

[Boellaard et al., 2015] Boellaard, R., Delgado-Bolton, R., Oyen, W. J. G., Giammarile, F., Tatsch, K., Eschner, W., Verzijlbergen, F. J., Barrington, S. F., Pike, L. C., Weber, W. A., Stroobants, S., Delbeke, D., Donohoe, K. J., Holbrook, S., Graham, M. M., Testanera, G., Hoekstra, O. S., Zijlstra, J., Visser, E., Hoekstra, C. J., Pruim, J., Willemsen, A., Arends, B., Kotzerke, J., Bockisch, A., Beyer, T., Chiti, A., and Krause, B. J. (2015). FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *European Journal of Nuclear Medicine and Molecular Imaging*, 42(2):328–354.

[Boellaard et al., 2001] Boellaard, R., van Lingen, A., and Lammertsma, A. A. (2001). Experimental and Clinical Evaluation of Iterative Reconstruction (OSEM) in Dynamic PET: Quantitative Characteristics and Effects on Kinetic Modeling. *Journal of Nuclear Medicine*, 42(5):808–817.

[Button et al., 2013] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365–376.

[Cannon et al., 2006] Cannon, D. M., Ichise, M., Fromm, S. J., Nugent, A. C., Rollis, D., Gandhi, S. K., Klaver, J. M., Charney, D. S., Manji, H. K., and Drevets, W. C. (2006). Serotonin Transporter Binding in Bipolar Disorder Assessed using [11C]DASB and Positron Emission Tomography. *Biological Psychiatry*, 60(3):207–217.

[Carp, 2012a] Carp, J. (2012a). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, 6:149.

[Carp, 2012b] Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1):289–300.

[Chen et al., 2014] Chen, G., Adleman, N. E., and Saad, Z. S. (2014). Applications of multivariate modeling to neuroimaging group analysis: A comprehensive alternative to univariate general linear model. *NeuroImage*, 99:571–588.

[Chen et al., 2018] Chen, K. T., Salcedo, S., Chonde, D. B., Izquierdo-Garcia, D., Levine, M. A., Price, J. C., Dickerson, B. C., and Catana, C. (2018). MR-assisted PET motion correction in simultaneous PET/MRI studies of dementia subjects. *Journal of Magnetic Resonance Imaging*, 48(5):1288–1296.

[Churchill et al., 2012] Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J. E., Graham, S. J., and Strother, S. C. (2012). Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Human Brain Mapping*, 33(3):609–627.

[Churchill et al., 2015] Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., and Strother, S. C. (2015). An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. *PLoS ONE*, 10(7):1–25.

[Cohen and Klunk, 2014] Cohen, A. D. and Klunk, W. E. (2014). Early detection of Alzheimer's disease using PiB and FDG PET. *Neurobiology of Disease*, 72:117–122. Special Issue: Metabolic disorders and neurodegeneration.

[Comley et al., 2013] Comley, R. A., Salinas, C. A., Slifstein, M., Petrone, M., Marzano, C., Bennacef, I., Shotbolt, P., Van der Aart, J., Neve, M., Iavarone, L., Gomeni, R., Laruelle, M., Gray, F. A., Gunn, R. N., and Rabiner, E. A. (2013). Monoamine transporter occupancy of a novel triple reuptake inhibitor in baboons and humans using positron emission tomography. *The Journal of Pharmacology and Experimental Therapeutics*, 346(2):311–7.

[Danad et al., 2014] Danad, I., Uusitalo, V., Kero, T., Saraste, A., Raijmakers, P. G., Lammertsma, A. A., Heymans, M. W., Kajander, S. A., Pietila, M., James, S., Sorensen, J., Knaapen, P., and Knuuti, J. (2014). Quantitative Assessment of Myocardial Perfusion in the Detection of Significant Coronary Artery Disease: Cutoff Values and Diagnostic Accuracy of Quantitative [15O]H2O PET Imaging. *Journal of the American College of Cardiology*, 64(14):1464–1475.

[Eklund et al., 2016] Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905.

[Fischer et al., 2009] Fischer, B., Lassen, U., Mortensen, J., Larsen, S., Loft, A., Bertelsen, A., Ravn, J., Clementsen, P., Høgholm, A., Larsen, K., Rasmussen, T., Keiding, S., Dirksen, A., Gerke, O., Skov, B., Steffensen, I., Hansen, H., Vilmann, P., Jacobsen, G., Backer, V., Maltbæk, N., Pedersen, J., Madsen, H., Nielsen, H., and Højgaard, L. (2009). Preoperative Staging of Lung Cancer with Combined PET-CT. *New England Journal of Medicine*, 361(1):32–39.

[Fischl et al., 2004] Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., and Dale, A. M. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11–22.

[Fisher et al., 2017] Fisher, P. M., Ozenne, B., Svarer, C., Adamsen, D., Lehel, S., Baaré, W. F. C., Jensen, P. S., and Knudsen, G. M. (2017). BDNF val66met association with serotonin transporter binding in healthy humans. *Translational Psychiatry*, 7(2):e1029.

[Forman et al., 2011] Forman, C., Aksoy, M., Hornegger, J., and Bammer, R. (2011). Self-encoded marker for optical prospective head motion correction in MRI. *Medical Image Analysis*, 15(5):708–719.

[Frankle et al., 2004] Frankle, W. G., Huang, Y., Hwang, D.-R., Talbot, P. S., Slifstein, M., Van Heertum, R., Abi-Dargham, A., and Laruelle, M. (2004). Comparative evaluation of serotonin transporter radioligands 11C-DASB and 11C-McN 5652 in healthy humans. *Journal of Nuclear Medicine*, 45(4):682–694.

[Freire and Mangin, 2001] Freire, L. and Mangin, J.-F. (2001). Motion Correction Algorithms May Create Spurious Brain Activations in the Absence of Subject Motion. *NeuroImage*, 14(3):709–722.

[Frick et al., 2015] Frick, A., Åhs, F., Engman, J., Jonasson, M., Alaie, I., Björkstrand, J., Frans, Ö., Faria, V., Linnman, C., Appel, L., Wahlstedt, K., Lubberink, M., Fredrikson, M., and Furmark, T. (2015). Serotonin Synthesis and Reuptake in Social Anxiety Disorder. *JAMA Psychiatry*, 72(8):794–802.

[Friston et al., 1995] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210.

[Frokjaer et al., 2015] Frokjaer, V. G., Pinborg, A., Holst, K. K., Overgaard, A., Henningsson, S., Heede, M., Larsen, E. C., Jensen, P. S., Agn, M., Nielsen, A. P., Stenbaek, D. S., Da Cunha-Bang, S., Lehel, S., Siebner, H. R., Mikkelsen, J. D., Svarer, C., and Knudsen, G. M. (2015). Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: A positron emission tomography study. *Biological Psychiatry*, 78(8):534–543.

[Gabrieli et al., 2015] Gabrieli, J. D., Ghosh, S. S., and Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, 85(1):11–26.

[Gelman and Geurts, 2014] Gelman, A. and Geurts, H. M. (2014). The statistical crisis in science. *American Scientist*, 102:460–65.

[Ginovart et al., 2001] 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*, 21(11):1342–1353.

[Golland and Fischl, 2003] Golland, P. and Fischl, B. (2003). Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies. In Taylor, C. and Noble, J. A., editors, *Information Processing in Medical Imaging*, pages 330–341, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Goodman et al., 2016] Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12.

[Green et al., 1994] Green, M. V., Seidel, J., Stein, S. D., Tedder, T. E., Kempner, K. M., Kertzman, C., and Zeffiro, T. A. (1994). Head Movement in Normal Subjects During Simulated PET Brain Imaging with and without Head Restraint. *Journal of Nuclear Medicine*, 35(9):1538–1546.

[Greve and Fischl, 2009] Greve, D. N. and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72.

[Greve and Fischl, 2017] Greve, D. N. and Fischl, B. (2017). False positive rates in surface-based anatomical analysis. *NeuroImage*, 171:6–14.

[Greve et al., 2016] Greve, D. N., Salat, D. H., Bowen, S. L., Izquierdo-Garcia, D., Schultz, A. P., Catana, C., Becker, J. A., Svarer, C., Knudsen, G. M., Sperling, R. A., and Johnson, K. A. (2016). Different partial volume correction methods lead to different conclusions: An 18F-FDG-PET study of aging. *NeuroImage*, 132:334–343.

[Greve et al., 2014] Greve, D. N., Svarer, C., Fisher, P. M., Feng, L., Hansen, A. E., Baare, W., Rosen, B., Fischl, B., and Knudsen, G. M. (2014). Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data. *NeuroImage*, 92:225–236.

[Hammoud et al., 2010] Hammoud, D. A., Endres, C. J., Hammond, E., Uzuner, O., Brown, A., Nath, A., Kaplin, A. I., and Pomper, M. G. (2010). Imaging serotonergic transmission with [11C]DASB-PET in depressed and non-depressed patients infected with HIV. *NeuroImage*, 49(3):2588–2595.

[Hansen et al., 1999] Hansen, L., Larsen, J., Nielsen, F., Strother, S., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., and Paulson, O. (1999). Generalizable patterns in neuroimaging: How many principal components? *NeuroImage*, 9(5):534–544.

[Hinderberger et al., 2016] Hinderberger, P., Rullmann, M., Drabe, M., Luthardt, J., Becker, G. A., Blüher, M., Regenthal, R., Sabri, O., and Hesse, S. (2016). The effect of serum BDNF levels on central serotonin transporter availability in obese versus non-obese adults: A [11C]DASB positron emission tomography study. *Neuropharmacology*, 110:530–536.

[Hirvonen et al., 2015] Hirvonen, J., Tuominen, L., Någren, K., and Hietala, J. (2015). Neuroticism and serotonin 5-HT1A receptors in healthy subjects. *Psychiatry Research*, 234(1):1–6.

[Holm et al., 1995] Holm, S., Toft, P., and Jensen, M. (1995). Estimation of the noise contributions from blank, transmission and emission scans in PET. In *1995 IEEE Nuclear Science Symposium and Medical Imaging Conference Record*, volume 3, pages 1470–1474 vol.3.

[Hornboll et al., 2018] Hornboll, B., Macoveanu, J., Nejad, A., Rowe, J., Elliott, R., Knudsen, G. M., Siebner, H. R., and Paulson, O. B. (2018). Neuroticism predicts the impact of serotonin challenges on fear processing in subgenual anterior cingulate cortex. *Scientific Reports*, 8(1):17889.

[Houle et al., 2000] Houle, S., Ginovart, N., Hussey, D., Meyer, J. H., and Wilson, A. A. (2000). Imaging the serotonin transporter with positron emission tomography: Initial human studies with [11C]DAPP and [11C]DASB. *European Journal of Nuclear Medicine*, 27(11):1719–1722.

[Ichise et al., 2003] Ichise, M., Liow, J. S., Lu, J. Q., Takano, A., Model, K., Toyama, H., Suhara, T., Suzuki, K., Innis, R. B., and Carson, R. E. (2003). Linearized reference tissue parametric imaging methods: Application to [11C]DASB positron emission tomography studies of the serotonin transporter in human brain. *Journal of Cerebral Blood Flow and Metabolism*, 23(9):1096–1112.

[Innis et al., 2007] Innis, R. B., Cunningham, V. J., Delforge, J., Fujita, M., Gjedde, A., Gunn, R. N., Holden, J., Houle, S., Huang, S.-C., Ichise, M., Iida, H., Ito, H., Kimura, Y., Koeppe, R. A., Knudsen, G. M., Knuuti, J., Lammertsma, A. A., Laruelle, M., Logan, J., Maguire, R. P., Mintun, M. A., Morris, E. D., Parsey, R., Price, J. C., Slifstein, M., Sossi, V., Suhara, T., Votaw, J. R., Wong, D. F., and Carson, R. E. (2007). Consensus nomenclature for in vivo imaging of reversibly binding radioligands. *Journal of Cerebral Blood Flow & Metabolism*, 27(9):1533–1539.

[Ioannidis, 2005] Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124.

[Jin et al., 2013] Jin, X., Mulnix, T., Gallezot, J. D., and Carson, R. E. (2013). Evaluation of motion correction methods in human brain PET imaging-A simulation study based on human motion data. *Medical Physics*, 40(10):1–12.

[Jin et al., 2014] Jin, X., Mulnix, T., Sandiego, C. M., and Carson, R. E. (2014). Evaluation of Frame-Based and Event-by-Event Motion-Correction Methods for Awake Monkey Brain PET Imaging. *Journal of Nuclear Medicine*, 55(2):287–293.

[Jørgensen et al., 2018] Jørgensen, L., Weikop, P., Svarer, C., Feng, L., Keller, S., and Knudsen, G. (2018). Cerebral serotonin release correlates with [11C]AZ10419369 PET measures of 5-HT1B receptor binding in the pig brain. *Journal of Cerebral Blood Flow & Metabolism*, 38(7):1243–1252.

[Jovanovic et al., 2009] Jovanovic, H., Karlsson, P., Cerin, Å., Halldin, C., and Nordström, A.-l. (2009). 5-HT 1A receptor and 5-HTT binding during the menstrual cycle in healthy women examined with [11C] WAY100635 and [11C] MADAM PET. *Psychiatry Research: Neuroimaging*, 172(1):31–37.

[Kalbitzer et al., 2010] Kalbitzer, J., Erritzoe, D., Holst, K. K., Nielsen, F. A., Marner, L., Lehel, S., Arentzen, T., Jernigan, T. L., and Knudsen, G. M. (2010). Seasonal changes in brain serotonin transporter binding in short serotonin transporter linked polymorphic region-allele carriers but not in long-allele homozygotes. *Biological Psychiatry*, 67(11):1033–1039.

[Keller et al., 2013] Keller, S. H., Svarer, C., and Sibomana, M. (2013). Attenuation Correction for the HRRT PET-Scanner Using Transmission Scatter Correction and Total Variation Regularization. *IEEE Transactions on Medical Imaging*, 32(9):1611–1621.

[Kero et al., 2017] Kero, T., Nordström, J., Harms, H. J., Sörensen, J., Ahlström, H., and Lubberink, M. (2017). Quantitative myocardial blood flow imaging with integrated time-of-flight PET-MR. *EJNMMI Physics*, 4(1):1.

[Khohlmyer and Stearns, 2002] Khohlmyer, S. and Stearns, C. (2002). NEMA NU2-2001 performance results for the GE Advance PET system. *Record, 2002 IEEE*, 100(1):890–894.

[Kim et al., 2006] Kim, J. S., Ichise, M., Sangare, J., and Innis, R. B. (2006). PET Imaging of Serotonin Transporters with [11C]DASB: Test-Retest Reproducibility Using a Multilinear Reference Tissue Parametric Imaging Method. *Journal of Nuclear Medicine*, 47(2):208–214.

[Kober et al., 2012] Kober, T., Gruetter, R., and Krueger, G. (2012). Prospective and retrospective motion correction in diffusion magnetic resonance imaging of the human brain. *NeuroImage*, 59(1):389–398.

[Lamare et al., 2007] Lamare, F., Carbayo, M. J. L., Cresson, T., Kontaxakis, G., Santos, A., Rest, C. C. L., Reader, A. J., and Visvikis, D. (2007). List-mode-based reconstruction for respiratory motion correction in PET using non-rigid body transformations. *Physics in Medicine & Biology*, 52(17):5187–204.

[Lammertsma and Hume, 1996] Lammertsma, A. A. and Hume, S. P. (1996). Simplified Reference Tissue Model for PET Receptor Studies. *NeuroImage*, 158(4):153–158.

[Logan et al., 1996] Logan, J., Fowler, J. S., Volkow, D., Wang, G.-j., Ding, Y.-s., and Alexoff, D. L. (1996). Distribution Volume Ratios Without Blood

Sampling from Graphical Analysis of PET Data. *Journal of Cerebral Blood Flow and Metabolism*, 16(5):834–840.

[Mandeville et al., 2016] Mandeville, J. B., Sander, C. Y. M., Wey, H.-y., Hooker, J. M., Hansen, H. D., Svarer, C., Knudsen, G. M., and Rosen, B. R. (2016). A regularized full reference tissue model for PET neuroreceptor mapping. *NeuroImage*, 139:405–414.

[Matheson et al., 2015] Matheson, G. J., Schain, M., Almeida, R., Lundberg, J., Cselenyi, Z., Borg, J., Varrone, A., Farde, L., and Cervenka, S. (2015). Diurnal and seasonal variation of the brain serotonin system in healthy male subjects. *NeuroImage*, 112:225–231.

[Mc Mahon et al., 2016] Mc Mahon, B., Andersen, S. B., Madsen, M. K., Hjordt, L. V., Hageman, I., Dam, H., Svarer, C., da Cunha-Bang, S., Baare, W., Madsen, J., Hasholt, L., Holst, K., Frokjaer, V. G., and Knudsen, G. M. (2016). Seasonal difference in brain serotonin transporter binding predicts symptom severity in patients with seasonal affective disorder. *Brain*, 139(5):1605–1614.

[McCarthy et al., 2015] McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J.-A., Coman, I. L., and Kates, W. R. (2015). A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience*, 9:379.

[McMahon et al., 2018] McMahon, B., Nørgaard, M., Svarer, C., Andersen, S. B., Madsen, M. K., Baaré, W. F. C., Madsen, J., Frokjaer, V. G., and Knudsen, G. M. (2018). Seasonality-resilient individuals downregulate their cerebral 5-HT transporter binding in winter - A longitudinal combined 11C-DASB and 11C-SB207145 PET study. *European Neuropsychopharmacology*, 28(10):1151–1160.

[Meltzer et al., 1999] Meltzer, C. C., Kinahan, P. E., Greer, P. J., Nichols, T. E., Comtat, C., Cantwell, M. N., Lin, M. P., and Price, J. C. (1999). Comparative Evaluation of MR-based Partial-Volume Correction Schemes for PET. *Journal of Nuclear Medicine*, 40(12):2053–2065.

[Meyer et al., 2001] Meyer, J. H., Wilson, A. A., Ginovart, N., Goulding, V., Hussey, D., Hood, K., and Houle, S. (2001). Occupancy of serotonin transporters by paroxetine and citalopram during treatment of depression: A [C-

11]DASB PET imaging study. *American Journal of Psychiatry*, 158(11):1843–1849.

[Miller et al., 2016] Miller, J. M., Everett, B. A., Oquendo, M. A., Ogden, R. T., Mann, J. J., and Parsey, R. V. (2016). Positron emission tomography quantification of serotonin transporter binding in medication-free bipolar disorder. *Synapse*, 70(1):24–32.

[Montgomery et al., 2006] Montgomery, A. J., Thielemans, K., Mehta, M. A., Turkheimer, F., Mustafovic, S., and Grasby, P. M. (2006). Correction of head movement on pet studies: Comparison of methods. *Journal of Nuclear Medicine*, 47(12):1936–1944.

[Monti, 2011] Monti, M. M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in Human Neuroscience*, 5:28.

[Morch et al., 1997] Morch, N., Hansen, L., Strother, S., Svarer, C., Rottenberg, D., Lautrup, B., Savoy, R., and Paulson, O. (1997). Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. *Information Processing in Medical Imaging*, (1230):259–70.

[Morimoto et al., 2006] Morimoto, T., Ito, H., Takano, A., Ikoma, Y., Seki, C., Okauchi, T., Tanimoto, K., Ando, A., Shiraishi, T., Yamaya, T., and Suhara, T. (2006). Effects of image reconstruction algorithm on neurotransmission PET studies in humans: Comparison between filtered backprojection and ordered subsets expectation maximization. *Annals of Nuclear Medicine*, 20(3):237–243.

[Müller-Gärtner et al., 1992] Müller-Gärtner, H. W., Links, J. M., Prince, J. L., Bryan, R. N., McVeigh, E., Leal, J. P., Davatzikos, C., and Frost, J. J. (1992). Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. *Journal of Cerebral Blood Flow and Metabolism*, 12(4):571–583.

[Nichols and Holmes, 2002] Nichols, T. and Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25.

[Nogami et al., 2013] Nogami, T., Takano, H., Arakawa, R., Ichimiya, T., Fujiwara, H., Kimura, Y., Kodaka, F., Sasaki, T., Takahata, K., Suzuki, M., Nagashima, T., Mori, T., Shimada, H., Fukuda, H., Sekine, M., Tateno, A., Takahashi, H., Ito, H., Okubo, Y., and Suhara, T. (2013). Occupancy of serotonin and norepinephrine transporter by milnacipran in patients with major depressive disorder: a positron emission tomography study with [(11)C]DASB and (S,S)-[(18)F]FMeNER-D(2). *International Journal of Neuropsychopharmacology*, 16(5):937–943.

[Nørgaard et al., 2015] Nørgaard, M., Ganz, M., Fisher, P. M., Mahon, B. M., and Strother, S. C. (2015). Estimation of Regional Seasonal Variations in SERT-levels using the FreeSurfer PET pipeline: a reproducibility study. *MICCAI workshop on Computational Methods for Molecular Imaging 2015*, (October):1–12.

[Nørgaard et al., 2017] Nørgaard, M., Ganz, M., Svarer, C., Fisher, P. M., Churchill, N. W., Beliveau, V., Grady, C., Strother, S. C., and Knudsen, G. M. (2017). Brain Networks Implicated in Seasonal Affective Disorder: A Neuroimaging PET Study of the Serotonin Transporter. *Frontiers in Neuroscience*, 11:614.

[Ogawa et al., 2014] Ogawa, K., Tateno, A., Arakawa, R., Sakayori, T., Ikeda, Y., Suzuki, H., and Okubo, Y. (2014). Occupancy of serotonin transporter by tramadol: a positron emission tomography study with [11C]DASB. *International Journal of Neuropsychopharmacology*, 17(6):845–50.

[Ogden et al., 2007] Ogden, R. T., Ojha, A., Erlandsson, K., Oquendo, M. A., Mann, J. J., and Parsey, R. V. (2007). In vivo quantification of serotonin transporters using [11C]DASB and positron emission tomography in humans: modeling considerations. *Journal of Cerebral Blood Flow and Metabolism*, 27(1):205–217.

[Olesen et al., 2009] Olesen, O. V., Sibomana, M., Keller, S. H., Andersen, F., Jensen, J., Holm, S., Svarer, C., and Højgaard, L. (2009). Spatial resolution of the HRRT PET scanner using 3D-OSEM PSF reconstruction. *IEEE Nuclear Science Symposium Conference Record*, pages 3789–3790.

[Olesen et al., 2013] Olesen, O. V., Sullivan, J. M., Mulnix, T., Paulsen, R. R., Hojgaard, L., Roed, B., Carson, R. E., Morris, E. D., and Larsen, R. (2013). List-Mode PET Motion Correction Using Markerless Head Tracking: Proof-of-Concept With Scans of Human Subject. *IEEE Transactions on Medical Imaging*, 32(2):200–209.

[Ooi et al., 2009] Ooi, M. B., Krueger, S., Thomas, W. J., Swaminathan, S. V., and Brown, T. R. (2009). Prospective Real-Time Correction for Arbitrary Head Motion Using Active Markers. *Magnetic Resonance in Medicine*, 62(4):943–954.

[OpenScienceCollaboration, 2015] OpenScienceCollaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

[Orchard and Atkins, 2003] Orchard, J. and Atkins, M. (2003). Iterating registration and activation detection to overcome activation bias in fMRI motion estimates. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, 2879:886–893.

[Parsey et al., 2006a] Parsey, R. V., Hastings, R. S., Oquendo, M. A., Huang, Y.-y., Simpson, N., Arcement, J., Huang, Y., Ogden, R. T., Van Heertum, R. L., Arango, V., and Mann, J. J. (2006a). Lower Serotonin Transporter Binding Potential in the Human Brain During Major Depressive Episodes. *American Journal of Psychiatry*, 163(1):52–58.

[Parsey et al., 2006b] Parsey, R. V., Kent, J. M., Oquendo, M. A., Richards, M. C., Pratap, M., Cooper, T. B., Arango, V., and Mann, J. J. (2006b). Acute Occupancy of Brain Serotonin Transporter by Sertraline as Measured by [11C]DASB and Positron Emission Tomography. *Biological Psychiatry*, 59(9):821–828.

[Poldrack et al., 2017] Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J. B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126.

[Reyes et al., 2007] Reyes, M., Malandain, G., Koulibaly, P. M., Gonzalez-Ballester, M. A., and Darcourt, J. (2007). Model-based respiratory motion compensation for emission tomography image reconstruction. *Physics in Medicine & Biology*, 52(12):3579–600.

[Rousset et al., 1998] Rousset, O., Ma, Y., and Evans, A. (1998). Correction for partial volume effects in PET: Principle and validation. *Journal of Nuclear Medicine*, 39(5):904–911.

[Rousset et al., 2007] Rousset, O., Rahmim, A., Alavi, A., and Zaidi, H. (2007). Partial Volume Correction Strategies in PET. *PET Clinics*, 2(2):235–249.

[Sander et al., 2019] Sander, C. Y., Mandeville, J. B., Wey, H.-Y., Catana, C., Hooker, J. M., and Rosen, B. R. (2019). Effects of flow changes on radiotracer binding: Simultaneous measurement of neuroreceptor binding and cerebral blood flow modulation. *Journal of Cerebral Blood Flow & Metabolism*, 39(1):131–146.

[Schain et al., 2014] Schain, M., Varnäs, K., Cselényi, Z., Halldin, C., Farde, L., and Varrone, A. (2014). Evaluation of Two Automated Methods for PET Region of Interest Analysis. *Neuroinformatics*, 12(4):551–562.

[Schwarz et al., 2018] Schwarz, C. G., Gunter, J. L., Lowe, V. J., Weigand, S., Vemuri, P., Senjem, M. L., Petersen, R. C., Knopman, D. S., and Jack, C. R. (2018). A Comparison of Partial Volume Correction Techniques for Measuring Change in Serial Amyloid PET SUVR. *Journal of Alzheimer's Disease*, 67(1):181–195.

[Schwarz et al., 2017] Schwarz, C. G., Jones, D. T., Gunter, J. L., Lowe, V. J., Vemuri, P., Senjem, M. L., Petersen, R. C., Knopman, D. S., and Jack Jr, C. R. a. (2017). Contributions of imprecision in PET-MRI rigid registration to imprecision in amyloid PET SUVR measurements. *Human Brain Mapping*, 38(7):3323–3336.

[Simmons et al., 2011] Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366.

[Strother et al., 2004] Strother, S., Conte, S. L., Hansen, L. K., Anderson, J., Zhang, J., Pulapura, S., and Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *NeuroImage*, 23(Supplement 1):S196–S207.

[Strother et al., 2002] Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage*, 15(4):747–771.

[Studholme et al., 1999] Studholme, C., Hill, D., and Hawkes, D. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86.

[Sureau et al., 2008] Sureau, F. C., Reader, A. J., Comtat, C., Leroy, C., Ribeiro, M.-J., Buvat, I., and Trebossen, R. (2008). Impact of Image-Space Resolution Modeling for Studies with the High-Resolution Research Tomograph. *Journal of Nuclear Medicine*, 49(6):1000–1008.

[Tegeler et al., 1999] Tegeler, C., Strother, S. C., Anderson, J. R., and Kim, S.-G. (1999). Reproducibility of BOLD-based functional MRI obtained at 4 T. *Human Brain Mapping*, 7(4):267–283.

[Tuominen et al., 2017] Tuominen, L., Miettunen, J., Cannon, D. M., Drevets, W. C., Frokjaer, V. G., Hirvonen, J., Ichise, M., Jensen, P. S., Keltikangas-Jarvinen, L., Klaver, J. M., Knudsen, G. M., Takano, A., Suhara, T., and Hietala, J. (2017). Neuroticism Associates with Cerebral in Vivo Serotonin Transporter Binding Differently in Males and Females. *International Journal of Neuropsychopharmacology*, 20(12):963–970.

[Turkheimer et al., 2012] Turkheimer, F. E., Selvaraj, S., Hinz, R., Murthy, V., Bhagwagar, Z., Grasby, P., Howes, O., Rosso, L., and Bose, S. K. (2012). Quantification of ligand PET studies using a reference region with a displaceable fraction: application to occupancy studies with [11C]-DASB as an example. *Journal of Cerebral Blood Flow and Metabolism*, 32(1):70–80.

[Tyrer et al., 2016] Tyrer, A. E., Levitan, R. D., Houle, S., Wilson, A. A., Nobrega, J. N., Rusjan, P. M., and Meyer, J. H. (2016). Serotonin transporter binding is reduced in seasonal affective disorder following light therapy. *Acta Psychiatrica Scandinavica*, 134(5):410–419.

[van den Heuvel et al., 2003] van den Heuvel, O. A., Boellaard, R., Veltman, D. J., Mesina, C., and Lammertsma, A. A. (2003). Attenuation correction of PET activation studies in the presence of task-related motion. *NeuroImage*, 19(4):1501–1509.

[van der Kouwe et al., 2006] van der Kouwe, A. J. W., Benner, T., and Dale, A. M. (2006). Real-time rigid body motion correction and shimming using cloverleaf navigators. *Magnetic Resonance in Medicine*, 56(5):1019–1032.

[van Velden et al., 2009] van Velden, F. H., Kloet, R. W., van Berckel, B. N., Buijs, F. L., Luurtsema, G., Lammertsma, A. A., and Boellaard, R. (2009). HRRT Versus HR+ Human Brain PET Studies: An Interscanner Test-Retest Study. *Journal of Nuclear Medicine*, 50(5):693–702.

[Varoquaux et al., 2017] Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179.

[Whitley and Ball, 2002] Whitley, E. and Ball, J. (2002). Statistics review 4: sample size calculations. *Critical Care*, 6(4):335–341.

[Wienhard et al., 2002] Wienhard, K., Wienhard, K., Schmand, M., Schmand, M., Casey, M. E., Casey, M. E., Baker, K., Baker, K., Bao, J., Bao, J., Eriksson, L., Eriksson, L., Jones, W. F., Jones, W. F., Knoess, C., Knoess, C., Lenox, M., Lenox, M., Lercher, M., Lercher, M., Luk, P., Luk, P., Michel, C., Michel, C., Reed, J. H., Reed, J. H., Richerzhagen, N., Richerzhagen, N., Treffert, J., Treffert, J., Vollmar, S., Vollmar, S., Young, J. W., Young, J. W., Heiss, W. D., Heiss, W. D., Nutt, R., and Nutt, R. (2002). The ECAT HRRT: Performance and First Clinical Application of the New High Resolution Research Tomograph. *IEEE Transactions on Nuclear Science*, 49(1):104–110.

[Wittchen et al., 2011] Wittchen, H., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jonsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., van Os, J., Preisig, M., Salvador-Carulla, L., Simon, R., and Steinhausen, H.-C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(9):655–679.

[Yarkoni and Westfall, 2017] Yarkoni, T. and Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

[Zanderigo et al., 2017] Zanderigo, F., Mann, J. J., and Ogden, R. T. (2017). A hybrid deconvolution approach for estimation of in vivo non-displaceable binding for brain PET targets without a reference region. *PLoS One*, 12(5):1–29.

[Zientek et al., 2016] Zientek, F., Winter, K., Moller, A., Rullmann, M., Luthardt, J., Becker, G. A., Bresch, A., Patt, M., Sabri, O., Hilbert, A., and Hesse, S. (2016). Effortful control as a dimension of temperament is negatively associated with prefrontal serotonin transporter availability in obese and non-obese individuals. *European Journal of Neuroscience*, 44(7):2460–2466.

[Zwan et al., 2017] Zwan, M. D., Bouwman, F. H., Konijnenberg, E., van der Flier, W. M., Lammertsma, A. A., Verhey, F. R. J., Aalten, P., van Berckel, B. N. M., and Scheltens, P. (2017). Diagnostic impact of [18F]flutemetamol PET in early-onset dementia. *Alzheimer's Research & Therapy*, 9(1):2.

Appendix A

Paper [A]

Nørgaard M, Ganz M, Svarer C, Feng L, Ichise M, Lanzenberger R, Lubberink M, Parsey RV, Politis M, Rabiner EA, Slifstein M, Sossi V, Suhara T, Talbot PS, Turkheimer F, Strother SC, Knudsen GM. 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*, 2019 Feb;39(2):210-222. DOI: 10.1177/0271678X18770107.

Check for updates

Review Article

JCBFM

Journal of Cerebral Blood Flow &
Metabolism
0(00) 1–13
© Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0271678X18770107
journals.sagepub.com/home/jcbfm
SAGE

Cerebral serotonin transporter measurements with [^{11}C]DASB: A review on acquisition and preprocessing across 21 PET centres

Martin Nørgaard[1,2], Melanie Ganz[1,3], Claus Svarer[1], Ling Feng[1], Masanori Ichise[4], Rupert Lanzenberger[5], Mark Lubberink[6], Ramin V Parsey[7], Marios Politis[8], Eugenii A Rabiner[9,10], Mark Slifstein[7], Vesna Sossi[11], Tetsuya Suhara[4], Peter S Talbot[12], Federico Turkheimer[13], Stephen C Strother[14] and Gitte M Knudsen[1,2]

Abstract

Positron Emission Tomography (PET) imaging has become a prominent tool to capture the spatiotemporal distribution of neurotransmitters and receptors in the brain. The outcome of a PET study can, however, potentially be obscured by suboptimal and/or inconsistent choices made in complex processing pipelines required to reach a quantitative estimate of radioligand binding. Variations in subject selection, experimental design, data acquisition, preprocessing, and statistical analysis may lead to different outcomes and neurobiological interpretations. We here review the approaches used in 105 original research articles published by 21 different PET centres, using the tracer [^{11}C]DASB for quantification of cerebral serotonin transporter binding, as an exemplary case. We highlight and quantify the impact of the remarkable variety of ways in which researchers are currently conducting their studies, while implicitly expecting generalizable results across research groups. Our review provides evidence that the foundation for a given choice of a preprocessing pipeline seems to be an overlooked aspect in modern PET neuroscience. Furthermore, we believe that a thorough testing of pipeline performance is necessary to produce reproducible research outcomes, avoiding biased results and allowing for better understanding of human brain function.

Keywords

Positron Emission Tomography, preprocessing, kinetic modeling, [^{11}C]DASB, data sharing

Received 18 October 2017; Revised 7 March 2018; Accepted 7 March 2018

---

[1]Neurobiology Research Unit, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[2]Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
[3]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[4]Department of Functional Brain Imaging Research, National Institute of Radiological Sciences, National Institutes for Quantum and Radiological Science and Technology, Chiba, Japan
[5]Department of Psychiatry and Psychotherapy, Medical University of Vienna, Vienna, Austria
[6]Department of Nuclear Medicine and Positron Emission Tomography, Uppsala University, Uppsala, Sweden
[7]Department of Psychiatry, School of Medicine, Stony Brook University, Stony Brook, NY, USA
[8]Neurodegeneration Imaging Group, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK
[9]Imanova Limited, London, UK
[10]Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[11]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada
[12]Division of Neuroscience and Experimental Psychology, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK
[13]Department of Neuroimaging, King's College London, London, UK
[14]Rotman Research Institute at Baycrest, University of Toronto, Toronto, Canada

Corresponding author:
Gitte M Knudsen, Neurobiology Research Unit, Section 6931, Rigshospitalet 9, Blegdamsvej DK-2100, Copenhagen.
Email: gmk@nru.dk

Introduction

Positron Emission Tomography (PET) imaging with selective radiotracers has been extensively used as a tool for novel neuroscience research. PET neuroimaging often utilizes complex workflows, with multiple stages ranging from subject selection, experimental design, data acquisition, preprocessing, statistical analysis to the final neurobiological interpretation.

However, while most published articles utilizing molecular neuroimaging have mainly focused on extracting neuroscientifically relevant results, no articles have, to our knowledge, investigated the extent to which these findings may be significantly influenced by different sets of preprocessing steps (“preprocessing pipeline/stage”) applied while analyzing the data. A preprocessing pipeline in neuroimaging commonly refers to a set of steps used to denoise and remove artifacts in the data for subsequent statistical analysis (e.g. motion correction and outlier detection), thereby improving the overall quality of the data. However, choices made at any stage of a neuroimaging workflow may significantly affect the chosen steps in the preprocessing pipeline, limiting the generalizability of any preprocessing pipeline studied in isolation from a fixed neuroimaging workflow. For example, medical conditions preventing patients from staying still in the scanner (e.g. Parkinson’s Disease) may require more extensive correction of head movements in the preprocessing stage compared to healthy subjects.

Notably, to date, preprocessing developments in the PET neuroimaging community have often been focusing on an even more limited point of view than examining the overall preprocessing pipeline in isolation. The optimization of preprocessing steps typically entails only limited test data, and is often performed with the aim of optimizing only a single preprocessing step (e.g. kinetic modeling) without explicitly attempting to address potential interactions with other preprocessing steps, or with other stages of a given workflow. Examples of such potential confounds include: subject selection (e.g. of healthy versus diseased cohorts), differences in scanner resolution, duration of a scanning session, dynamic framing, injected dose/injected mass (data acquisition), differences in image reconstruction, motion correction, different kinetic modeling approaches used to estimate the availability of receptors/transporters (preprocessing), and different statistical model choices used to test for group or longitudinal differences (statistical analysis).

We here review and quantify the impact of the various data acquisition and preprocessing pipeline choices used to quantify the same biological target, using the serotonin transporter and the radioligand [^{11}C]DASB as exemplary case. We chose to specifically focus on the serotonin transporter using [^{11}C]DASB, because this is a well established radioligand in the field and has been used extensively to study various aspects of brain function ranging from schizophrenia to epilepsy (Figure 1).

Since [^{11}C]DASB was first described in 2000[1] through the end of March 2017, nearly 170 [^{11}C]DASB PET papers have been published, and this

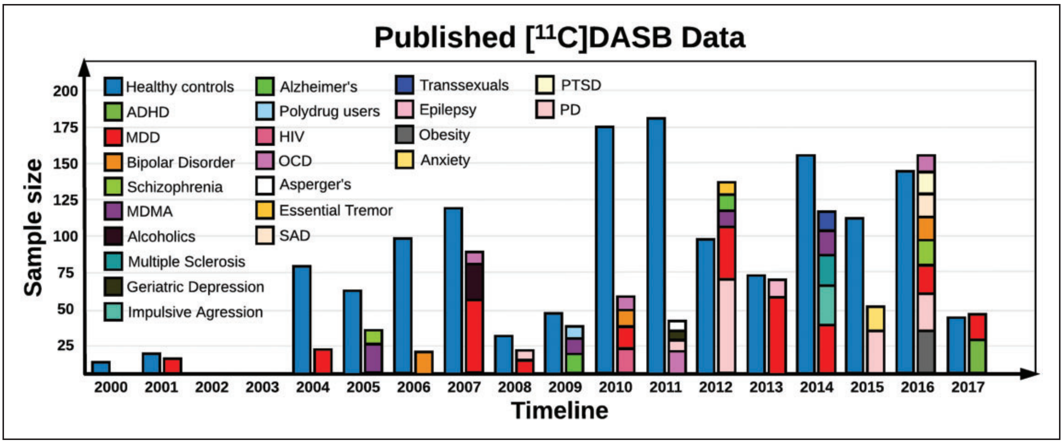

Figure 1. Timeline of number of patient and healthy controls in the 105 published [^{11}C]DASB studies. The colors indicate either healthy controls, or a specific disorder as a function of time and sample size. ADHD: attention-deficit/hyperactive disorder; MDD: major depressive disorder; MDMA: ecstasy; HIV: human immunodeficiency virus; OCD: obsessive compulsive disorder; SAD: seasonal affective disorder; PTSD: post-traumatic stress syndrome; PD: Parkinson’s disease.

number is growing by one to two articles per month. We systematically searched PubMed for studies using "[^{11}C]DASB and PET" in the time period between September 2000 to March 2017, and found a total of 169 publications. Non-human studies (N = 49), reviews (N = 4), and methodological papers (N = 12) were excluded due to substantial differences in acqusition and preprocessing, leaving 104 publications eligible for scrutiny. One paper not identified by the search,[2] was subsequently added, summing up to a total of 105 original research articles. We catalogued the different sample sizes and patient cohorts investigated in the published [^{11}C]DASB studies, the various data acquisition techniques used, and the preprocessing steps applied to the data. We systematically outline and quantify the impact of the remarkable variety of ways in which researchers are currently performing these studies, while implicitly expecting generalizable results across research groups. Although this review specifically focuses on the radioligand [^{11}C]DASB, the underlying considerations apply to any given PET or SPECT radiotracer, as optimal neuroimaging workflows are highly dependent on the inherent characteristics of the radioligand of interest.

Data acquisition workflow and outcome

In order to investigate the variability in data acquisition and preprocessing, we provide an overview of the different acquisition and preprocessing choices that have been made in previous studies. We also examine how differences in reported findings might be influenced by differences in methodologies. For this purpose, we extract the [^{11}C]DASB PET binding potentials (BP_{ND}) in striatum and anterior cingulate cortex (ACC) as well as other relevant information from 90 studies with healthy controls encompassing a total of 1856 healthy controls. We chose to examine the healthy controls only because they serve as null data, achieved with different experimental designs. The available BP_{ND}'s and standard deviations from the published studies were used as the dependent variable in separate linear models, correcting for the number of healthy controls included in the study, age, age standard deviation, choice of MRI hardware, choice of PET hardware, number of frames, injected dose, motion correction, choice of volumes-of-interest (VOI), and choice of kinetic modeling (Table S1). All covariates were standardized columnwise to have mean 0 and standard deviation 1. To limit the degrees of freedom, we did not specify any interactions in the linear model, despite their obvious existence (e.g. PET scanner × injected dose).

The omission of potential interactions is a limitation of the current analysis, but is driven by limited data.

Development of [^{11}C]DASB in PET neuroimaging and subject selection

N,N-dimethyl-2-(2-amino-4-cyanophenylthio)benzylamine, or more commonly referred to as DASB, was developed by Wilson et al. at the Center for Addiction and Mental Health, Toronto Canada, and their first-in-human study was published in 2000.[1,3]

Their preliminary analyses indicated that DASB radiolabeled with carbon-11 effectively penetrated the blood–brain barrier, and displayed retention characteristics in accordance with the known anatomical distribution of cerebral serotonin reuptake sites. In any aspect, [^{11}C]DASB turned out to be a highly suitable radiotracer to map the serotonin transporter using dynamic 4D PET imaging.

Since 2000, [^{11}C]DASB has been used extensively, so far by 21 PET centres, investigating various aspects of brain function. In Figure 1, we provide a timeline of the number of healthy controls and patient cohorts that have been investigated and published using [^{11}C]DASB. Whenever possible, we have attempted to correct the data in Figure 1 for duplicates, to encounter only the net number of included healthy volunteers from the [^{11}C]DASB PET studies.

Our analysis of the reported values from the literature suggests no statistical evidence for an impact of the number of subjects included in the study on BP_{ND} or between-subject variation.

We found a trend for an association between age and between-subject variation of ACC BP_{ND} ($P = 0.075$), suggesting that between-subject ACC BP_{ND} is more variable in elderly than in young controls. While this may be caused by cortical atrophy or other age-related disorders, it warrants further examination of how the impact of acquisition and preprocessing choices may vary as a function of age.

PET scanners and reconstructions

We found that in the 21 centres, 9 different scanners have been used (Figure 2). The first paper published by Houle et al.[3] (Center for Addiction and Mental Health, Toronto, Canada) presented data acquired with a Scanditronix/GEMS PC2048-15B 2D brain PET scanner, a state-of-the-art scanner from the late 80s. The data were attenuation corrected and reconstructed using filtered back-projection (FBP). The performance of the Scanditronix/GEMS PC2048-15B scanner was evaluated in 1989 by Holte et al.,[4] reporting the in-plane axial full-width half maximum (FWHM) to be 5.9 mm for direct planes, and 5 mm for cross planes in the central area of the field-of-view (FOV). In addition, with a coincidence timing window of 12.5 ns and a lower energy threshold of 300 keV,

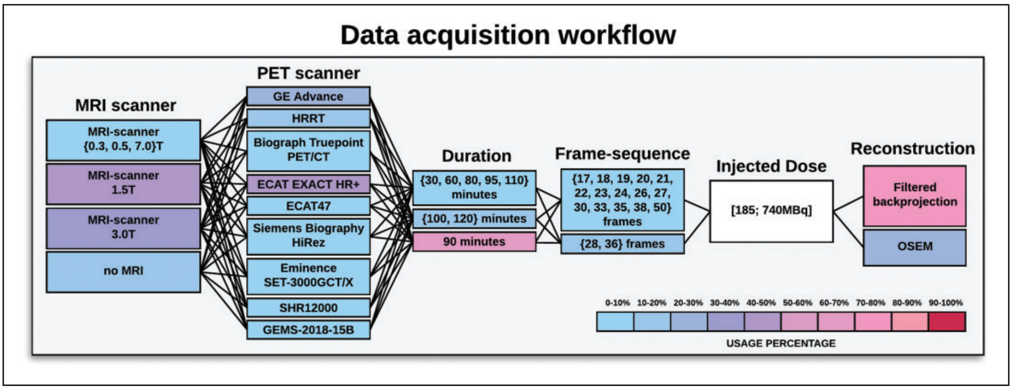

Figure 2. Schematic overview of the different data acquisition workflows used to acquire dynamic [^{11}C]DASB data. The workflow consists of scanners providing anatomical information, i.e. MRI scanners at various field strengths (Tesla), various PET scanners, duration of the dynamic PET acquisition, frame sequence used to temporally acquire 4D [^{11}C]DASB data, injected dose (ranging from approximately 100-740 MBq), and finally the reconstruction methods used to reconstruct the 4D PET sequence. The colors indicate the frequency per step that has been applied in a [^{11}C]DASB PET study out of the total 105 studies. Injected dose is filled as white, because it spans a continuous range and is highly subject-specific. The 4D imaging data are the output of the data acquisition workflow and input to the preprocessing workflow.

the average sensitivity (including 16% scatter) was 251 cps·MBq^{-1}·mL^{-1} for the direct planes, whereas the average sensitivity was 351 cps·MBq^{-1}·mL^{-1} for the cross planes. After the study by Houle et al.[3] and until 2008, a total of six DASB studies were conducted with the GEMS scanner, all published by the Center for Addiction and Mental Health, Toronto, Canada, including the first [^{11}C]DASB study discussing quantification strategies by Ginovart et al.[5]

After the first publication of [^{11}C]DASB, attention increased substantially around the World, motivating researchers to investigate new hypotheses related to the serotonin transporter. Consequently, resulting in a large number of different scanners used to map [^{11}C]DASB binding. Ogawa et al.[6] from Japan used an Eminence SET-3000GCT/X PET scanner (performance evaluated in 2006[7]) to investigate the effects of Tramodol for pain treatment; this is currently the only published [^{11}C]DASB study using this scanner. Another Japanese group[8] used an SHR12000 tomograph from Hamamatsu Photonics (performance evaluated in 2002[9]) to study the serotonin transporter in Alzheimer's Disease; this is the only [^{11}C]DASB study published using this scanner. Both of these scanners operate in 3D-mode, providing an excellent in-plane spatial resolution ranging from approximately 3 mm FWHM in the center of the FOV to 5 mm FWHM at 10 cm off center. This makes them somewhat ideal PET scanners to capture cortical features of the serotonin transporter, as on average, cortex is only 3 mm thick.[10]

The National Institute of Radiological Sciences in Chiba Japan, published two [^{11}C]DASB studies in 2006[11] and 2010[12]; these were the only studies using an ECAT47 PET scanner. This PET scanner also operates in 3D-mode, but unlike the Eminence and Hamamatsu scanners, which have an axial resolution of 3–5 mm FWHM, this scanner has an in-plane axial spatial resolution of 6.2 mm in the center of the FOV, and 7.2 mm at 10 cm off center.[13] This means that the spatial resolution is almost half as good in the center of the FOV, and more severe partial volume effects (PVEs) are to be expected. Several integrated PET/CT systems have also been used to map the serotonin transporter, including the Biograph HiRez[14] and the Biograph TruePoint,[15] both manufactured by Siemens, having a spatial resolution of approximately 4.5 mm. A total of seven published [^{11}C]DASB PET papers have used this scanner. The most commonly used PET scanners for measuring [^{11}C]DASB are the ECAT EXACT HR + PET scanner from Siemens (performance evaluated in 1997[16]), the GE Advance PET scanner from General Electric (performance evaluated in 2002[17]) and the High Resolution Research Tomography (HRRT) PET scanner from Siemens (performance evaluated in 2002[18]). Van Velden et al.[19] directly compared the performances of the HRRT and HR + scanner in 2009. The in-plane spatial resolution of the HRRT is 2.3–3.4 mm FWHM, whereas the in-plane spatial resolution of the HR + scanner is 4.3–8.3 mm FWHM.

Furthermore, the sensitivity of the HRRT is higher than that of the HR + scanner, 39.8 kcps·kBq^{-1}·mL^{-1} compared to 21.9 kcps·kBq^{-1}·mL^{-1}, respectively.

Finally, the GE advance PET scanner from General Electrics has an in-plane spatial resolution of 4.4 mm

FWHM in the center of the FOV, and 6 mm FWHM in the outer FOV.[17] The GE Advance scanner sensitivity is approximately 27.6 kcps·kBq^{-1}·mL^{-1}, placing it as the second best performing scanner regarding scanner sensitivity, with the HRRT in first place, and the ECAT EXACT HR + in third place.

Our analysis of the reported values from the papers revealed that BP_{ND} in both striatum and ACC was associated with the PET-scanner used in the study. This was also true for the between-subject standard deviation of BP_{ND} (Figure S1). A higher scanner resolution was associated with higher BP_{ND}'s and higher between-subject standard deviations ($P = 0.027$).

This means that more subjects are needed to detect a statistical difference in a group analysis.

On the other hand, the larger between-subject variability may also be caused by increased ability to detect subject-specific binding, as reflected by a higher resolution scanner.

The HRRT scanner has high sensitivity, but is limited by relatively small detector elements which means, that the number of acquired counts is lower than in other scanners, potentially resulting in more noisy data.

Moreover, the spatial resolution differs significantly between scanners with the HR + being nearly isotropic, whereas the GE Advance has a much better axial resolution than transaxial resolution (non-isotropic voxels). This means that the resolution is dependent on the orientation of the image, resulting in different spill-over effects of the tracer in different directions. This makes it difficult to correct for PVEs, and may consequently interact with subsequent preprocessing steps such as motion correction, co-registration and normalization to a standard space.

Instituting a more standardized policy for the reporting/usage of PET scanner performances should ensure that future readers are better able to effectively evaluate and understand the potential biases and uncertainties of the data. We note that researchers often only report the FWHM in the center of the FOV when publishing papers, creating a limited/biased interpretation if cortical regions are the primary region of interest.

In addition, reviewers should pay special attention to the use of 2D reconstruction over 3D reconstruction, non-isotropic over isotropic resolution, and if any additional smoothing steps are applied to the data (e.g. Strecker et al.[20]), as these steps significantly degrade the spatial resolution.

Anatomical information from magnetic resonance imaging

Several different techniques have been used to provide the anatomical information needed to guide the functional information provided by the PET data. The most common procedure is to acquire an anatomical T1-weighted Magnetic Resonance Image (MRI) as a reference image and spatially align the two images (co-registration). However, the field-strength of the MRI scanner will have an impact on the reconstructed MRI image, affecting both the subsequent parcellation of the brain into anatomical subregions and the co-registration to the PET data. Tradeoffs between spatial and temporal resolution and signal/noise also matter, but this topic is considered beyond the scope of this paper.

In the reported [^{11}C]DASB studies, the field-strength used for the MRI scanners includes 0.3T,[8] 0.5T,[21] 1.5T,[22] 3.0T[23] to 7.0T.[24] When no MRI is acquired, the PET image is most often either normalized to a common atlas space (e.g. Lanzenberger et al.[25]), in which generic regions have been predefined, or manual delineations are applied directly on the PET image. This requires additional smoothing or resampling steps and interacts with nearly all data acquisition steps such as the resolution of the PET scanner, duration, framing and injected dose (Figure 2, described in detail below). In addition, because a standard MRI atlas does not follow the subject-specific anatomy (i.e. cortical folding patterns), it is likely that this procedure will exacerbate the PVE when evaluating regional PET distributions compared to when a subject-specific MRI is available.

For example, if the PET-MRI co-registration is inaccurate, the PET signal might seem to originate from white-matter signal instead of gray matter (GM), or vice versa. In total, we found six different methods whereby anatomical information has been extracted. In our analysis, we found no evidence for an impact of choice of MRI-scanner on BP_{ND} or between-subject variation.

The two to date most widely published methods are acquisitions of either a 1.5T (43%) or 3.0T (32%) T1-weighted MRI (Figure 2). Not unexpectedly, the more recent publications tend to use 3T MRI scanners because most institutes regularly update their MR scanners, with newer ones having higher field strength. However, the extent to which differences in MRI acquisition might affect the final outcome of a complex workflow is largely unknown.

Data acquisition (duration, framing, injected dose, reconstruction)

Variations in [^{11}C]DASB PET data acquisition are distributed across a parameter space containing the (1) time duration of the scanning session, (2) dynamic framing (time-sequence), (3) injected dose and (4) PET reconstruction. Unless list mode acquisitions are available, dynamic PET studies are mostly acquired for a fixed time duration, with multiple 3D-frame acquisitions distributed over a pre-defined time period. However, the chosen framing varies substantially from study to study, and a total of 17 different sequences have been used so far, i.e. framing $\in \{17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 30, 33, 35, 36, 38, 50\}$ frames. This choice will affect the signal distribution within the acquisition space (FOV), since reduced frame length will result in a reduction in true counts per frame, especially in late frames where the radioactive tracer has decayed due to the half-life of ^{11}C (i.e. $\lambda_{1/2} = 20.3$ min). We identified a positive association between number of frames and striatal BP_{ND} and ACC BP_{ND} but closer inspection revealed that this observation was driven mainly by the HRRT-scanner settings from Denmark using 36 frames and the high-resolution SHR12000-scanner from Japan using 38 frames. As high-resolution scanners increase the binding, this effect may be at least partially explained as a scanner × frames interaction.

The scan duration of these dynamic PET studies also varies substantially, ranging from 30 min to 120 min, i.e. duration $\in \{30, 60, 80, 90, 95, 100, 110, 120\}$ minutes. Although, Ogden et al.[26] argued that 100 mins of scanning time was sufficient, this recommendation has not been followed in all subsequent studies, with most studies choosing 90 min of acquisition time (Figure 2). The required duration of scanning time, however, depends on the neuroscientific question, as various brain structures will have different uptake dynamics.[27]

The injected dose also varies substantially between approximately 100 MBq to 740 MBq, and the dose varies substantially not only between studies, but also between subjects (Table S1).

All studies reported high molar radioactivity; however, it has never been formally established in a test–retest study with substantially different doses, if a higher injected mass (or mass/kg body weight) leads to a reduction in cerebral [^{11}C]DASB binding. In a large sample of 108 individuals from our own group, we found no evidence for an association between global [^{11}C]DASB binding and injected mass/kg (McMahon et al. 2017, unpublished work). In our analysis, we found no evidence for an impact of injected dose on BP_{ND} or between-subject variability.

Depending on the scanner sensitivity, the injected dose can impact signal-to-noise ratio (SNR). Information about, e.g., range of average counts per minute per study/subject could be interesting information to have access to for analysis, as we did not have access to the individual injected doses, but rather the within-study average injected doses.

The reconstruction of the PET images from the scanner has also been differently performed. Morimoto et al.[28] compared [^{11}C]DASB binding in seven healthy subjects using images reconstructed with either a filtered backprojection (FBP) algorithm or the ordered subsets expectation maximization (OSEM) algorithm.

The study by Morimoto et al. was executed using the data acquisition workflow parameters: ECAT EXACT HR+, 1.5T MRI, 90-min dynamic PET acquisition in 2D-mode, 27 frames, injected dose of 170.2 ± 56.1 MBq. While there have been several reports suggesting a small bias using some versions of OSEM,[29,30] Morimoto et al. reported no statistically significant differences in any regions between images reconstructed with FBP and OSEM, suggesting that these two algorithms may be used interchangeably in the reconstruction of 4D PET data. Certain PET scanners (e.g. the HRRT) do not allow for a direct use of FBP due to the inherent geometry of the scanner, thereby restricting image reconstruction to the use of iterative reconstruction techniques such as the OSEM algorithm. However, techniques have been developed that allow 3D-FBP on the HRRT, but due to poor noise performance they are not widely used. To summarize the data acquisition workflow section, the most widely published workflow consists of: 1.5T MRI (43%), ECAT EXACT HR+(43%), 90-min acquisition (65%), 26 frames (17%), and FBP to reconstruct the 4D PET data (72%).

The "preprocessing pipeline" for [^{11}C]DASB PET quantification

Motion correction

Motion correction (MC) algorithms for dynamic PET studies have been developed to remove inherent motion artefacts from the data. The most popular head MC technique is between-frame-correction where either all or a subset of the remaining images are registered to a chosen reference image. Of the 105 studies, 43 studies (41%) leave out any type of MC, arguing that fixing the subject in the scanner using, e.g., a thermoplastic mask sufficiently limits motion. Twenty-nine studies used between-frame-correction to correct for motion without explicitly specifying the exact procedure (e.g. James et al.[31]). Twenty-one studies used between-frame-correction to correct for motion, by aligning all frames to a frame with high SNR (e.g. Frokjaer et al.[32]). Ten studies used either a mean or a summed PET image over all frames to correct for motion (e.g. Cannon et al. 2007[33]), and two studies used either a partially summed image[34] or a reference frame[35] to perform between-frame MC, but only frames where the researcher observed motion are aligned, leaving the frames without motion untouched.

The latter method not only introduces a user-dependent bias, it also raises the question: given that motion is present in the data, how much movement is needed in order to perform MC?

Overall, this results in five different ways in which MC has been applied/not applied in the [^{11}C]DASB literature. In our analysis, we grouped the analysis into motion versus no motion and observed a trend for significance ($P = 0.064$) of the use of MC and striatal between-subject variability, suggesting that MC lowers between-subject variability in the striatum with 0.035 compared to without MC (Figure S2). This translates into 26% fewer subjects needed in a group analysis to obtain similarly powered statistical tests (see calculation in supplementary).

MC in the absence of motion will lead to some degree of smoothing, which may to some extent account for the observed reduced between-subject variability. In addition, potential effects of motion within a frame are often neglected, even though several solutions have been suggested such as MOLAR[75] (Motion-Compensation OSEM List-mode Algorithm for Resolution-Recovery Reconstruction) or Tracoline[76] (List-mode PET MC using markerless head tracking), given that list-mode data are available.

MC is often carried out using different software packages (AIR, FSL and SPM), which all have different implementation and precision of similar methods but based on different cost functions. To our knowledge, the effect of various software packages on MC performance has not yet been investigated in dynamic 4D PET imaging. In addition, frame-by-frame MC without re-doing the image reconstruction may result in errors in attenuation correction, which is often neglected (Van den Heuvel et al.[73]).

Co-registration

Accurate co-registration of PET and MR images is an important step, not the least when PET Partial Volume Correction (PVC) and parcellation of regions are carried out, when integrating multimodal neuroimaging data.[36] Ninety-eight percent of all studies used a normalized mutual information (NMI) registration algorithm to perform the co-registration but the explicit procedure differs across studies that use various software packages, including FSL and SPM. Each co-registration technique is based on a cost-function, aiming to minimize the registration error (e.g. sum of least-squares or mutual information) of the two datasets being aligned (MRI and PET). This cost-function is often based on shared information between the two datasets being aligned (e.g. cortical boundaries), making them somewhat dependent on the intensity distribution and resolution of the acquired data. The remaining 2%[37,38] used a boundary-based registration (BBR) algorithm to co-register the T1-weighted MRI with the PET image. BBR also contains a mutual information component, but puts an additional cost on the cortical boundaries being aligned. The co-registration preprocessing step potentially depends on the spatial- and temporal distribution of the PET signal, and will therefore be sensitive to the chosen cost-function. For example, the serotonin transporter is only modestly expressed in the neocortex, and the boundary-based algorithm may therefore not be the optimal registration algorithm to capture cortical folding patterns, particularly not if the PET scanner resolution is limited. In addition, brain areas located in close vicinity to ventricles and cerebrospinal fluid (CSF) will suffer more from PVEs, depending on the resolution of the PET scanner and the radiotracer being used, especially when data with non-isotropic spatial resolution were acquired.

Delineation of volumes of interest

Many neuroimaging experiments are based on hypotheses relating to specific anatomical brain regions, often referred to as VOIs. As mentioned previously, for PET, this generally requires co-registration with a structural MRI scan with anatomical labels. However, there is currently no consensus in the [^{11}C]DASB PET community about which atlas generates the best set of VOIs. Whereas a single study used the probabilistic Harvard-Oxford atlas to delineate VOIs,[31] 14 published papers used PVElab, which is a data-driven anatomical probability-based labeling approach based on MRI templates from 10 healthy volunteers (e.g. Frokjaer et al.[39]). Nine studies used the Desikan/Killiany atlas (e.g. from FreeSurfer) which involves a data-driven technique, providing the researcher with a subject-specific anatomical labeling, given that they have acquired a subject-specific T1-weighted MRI (e.g. Ganz et al.[37]). Seven studies used the anatomical automatic labeling (AAL) atlas offered by the SPM software (e.g. Savli et al.[40]). The AAL atlas does not provide unique subject-specific anatomical labeling, but can be used for group analyses, where all subject-specific PET scans have been normalized to AAL standard space. Seven studies used the Hammers atlas, which is a probabilistic brain atlas based on 83 manually delineated regions drawn on MR images of 30 healthy subjects in native space, subsequently spatially normalized to a standard brain from the Montreal Neurological Institute (MNI) (e.g. Hinz et al.[41]). Fourteen studies used an atlas-based procedure, without explicitly stating the exact labeling approach, mostly being based on local procedures and study-specific atlases (e.g. Takano et al.[42]). These atlases are often based on data obtained from a set of young and healthy subjects, in which manually delineated regions have been drawn in native subject-space prior to spatial normalization to MNI-space. Ten published studies used an "automatic method" to obtain VOIs, stating that the anatomical labeling was unbiased with respect to any user interactions (e.g. Tyrer et al.[14]).

Somewhat surprisingly, 38% of all published [11C]DASB studies included in this review, manually define their own VOIs, also in some recent studies.[43,44]

In our analysis, we found a striatal BP_{ND} x VOI interaction, suggesting that some definitions of volumes produce either higher or lower BP_{ND} compared to others (Figure S3). Since this step may interact with all previous steps, we are cautious to make any firm conclusions based on this.

Hammers atlas and manual delineations contributed to the most variation, but should ideally also be split into additional sub-categories depending on the operational criteria, and whether the delineation was performed in PET or MRI space (Table S1). In addition, it is expected that the variability will increase as the size of the VOI decreases, but with limited reports on size of VOIs in the published studies, this reduces our ability to assess the impact of atlas choice. Nevertheless, what we can conclude is that the choice of atlas can produce widely different outcomes, as highlighted in Table S1.

Even though manual anatomical labeling seems to be the most popular, it may impose an interrater variability/bias in the subsequent data analysis and interpretation, unless well-defined operational criteria and blindness to subject diagnosis are applied.

Another potential issue with both manual delineation and atlases is that even though the tracer distribution within an anatomical VOI is assumed to be homogeneous, this is often not the case and accordingly, structural homogeneously defined VOIs may therefore misrepresent the radioligand concentration within that region (e.g. the thalamus). Correct anatomical labeling is critically important in many dynamic PET studies, because the PET data suffers significantly from PVEs.

We recommend that researchers provide explicit specifications about VOI definitions in the supplementary material, and if possible, attach the 3D anatomical labelings in appropriate formats. This is an approach also supported by researchers in the fMRI field.[45]

Partial volume correction

In PET studies, it can be difficult to assess the extent to which an observed difference in PET signal is caused by a change in the imaging target distribution, if it is due to less GM, or if it is due to limited PET scanner resolution causing the PET signal to spill in or out of relatively homogeneous tissue regions. Partial Volume Correction (PVC) is not commonly used in [11C]DASB PET imaging (Figure 3). Only four published [11C]DASB studies have used Muller-Gartner PVC, to correct for PVEs.[46–48,74]

If there is little evidence for differences in brain volumes, the application of PVC techniques may lead to noise amplification, and extreme care should therefore be taken when interpreting the results.[36] In addition, PVC is MR scanner and sequence dependent due to variability of segmentation results from the MRI. For an in-depth discussion of PVC techniques in PET imaging, we refer the reader to the paper by Erlandsson et al.[49]

Quantification of [11C]DASB PET data

The final step in processing of [11C]DASB PET data is kinetic modeling which is applied to the preprocessed 4D PET data. All the kinetic modeling approaches used for quantification of [11C]DASB PET data are displayed in Figure 3, including the frequency of their use. The quantification of tracer kinetics of the serotonin transporter in vivo has been applied extensively and in various formats, providing information about binding in specific VOIs. The gold standard is to obtain arterial blood samples in parallel with the dynamic PET scan, providing an arterial input function (AIF) for subsequent kinetic modeling.[5]

However, the use of arterial sampling requires invasive techniques, which often imposes additional discomfort to the subject being scanned. Furthermore, blood sample analysis (on-line vs. manual sampling, including frequency of sampling), metabolite estimation (HPLC or fraction-collector) and interpolation (fitting a power function) can add additional variation to the data analysis. Two-tissue compartment modeling (2TCM) with an AIF is considered state-of-the-art in the PET literature, but once validated, tissue reference methods may be used instead. The kinetic models used in [11C]DASB neuroimaging include both reference tissue methods and methods with an AIF (Figure 3). Reference tissue methods obviate invasive arterial sampling, but they rely on the assumption and identification of a reference region with non-specific binding characteristics.

In the [11C]DASB literature, cerebellum (possibly excluding vermis) serves as a reference region, because it is considered to be devoid of serotonin transporters. However, there is currently no consensus among researchers about the validity of cerebellum as a reference region. Some researchers argue for[5,40,50] and others against, as DASB binding in the cerebellum has been shown to be displaced by SSRIs.[51–53] Even among the researchers using cerebellum as a reference region, there is no consensus about how exactly the reference region must be defined.[37,52,54] A recent

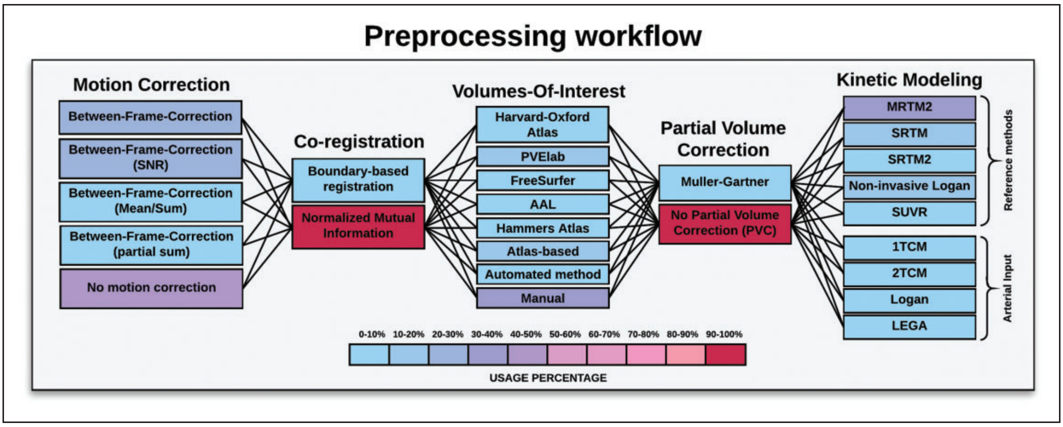

Figure 3. Schematic overview of the various preprocessing steps used in analyzing dynamic [^{11}C]DASB data. This ranges from different motion correction techniques, co-registration, volume-of-interest definitions, partial volume correction, and kinetic modeling. The colors indicate the percentage, in which a given step has been applied in the 105 [^{11}C]DASB PET studies.

investigation of cerebellar heterogeneity and its impact on PET data quantification of 5-HT receptor radioligands, based on a large sample of 100 [^{11}C]DASB HRRT scans, concluded that there are differences in radioactivity uptake between cerebellar subregions.[37]

New kinetic models are continually being developed and refined, and to date nine different approaches have been applied. Published studies that include blood sampling and use of an AIF over the last couple of years have become less common, with currently four different approaches used to perform the kinetic modeling of [^{11}C]DASB. Only three published studies have used a one-tissue compartment model (1TCM) with an AIF to capture the features of the serotonin transporter.[5,55,56] Eight studies used a 2TCM with an AIF (e.g. van de Giessen et al.[57]), and eight studies have used the Logan method with an AIF (e.g. Murthy et al.[58]). Finally, the likelihood estimation graphical analysis (LEGA) method (maximum likelihood estimation of the Logan) using an AIF has been used in eight published studies, including one of the three test–retest studies that evaluate reproducibility of [^{11}C]DASB,[26] as discussed in more detail below.

Forty-four published studies (38% in total) have used the multilinear reference tissue model 2 (MRTM2), developed by Ichise et al.,[59] to quantify tracer kinetics of [^{11}C]DASB (e.g. Fisher et al.[60]). Twelve studies used the simplified reference tissue model (SRTM) developed by Lammertsma and Hume in 1996,[61] and eight studies a constrained version of the same model, SRTM2.[62]

The non-invasive Logan method is used in 22 published studies.[63] Finally, four studies have used the ratio of standardized uptake values (SUVR) defined by the SUV of a given VOI to the SUV of a reference VOI (i.e., the cerebellum has been used as a reference for DASB binding). When using SUVR as a direct measure for binding, the arterial input concentration is assumed to have a consistent shape between studies/subjects, and the area under the arterial input curve is assumed to be proportional to the injected dose/kg body weight.[64] This assumption applies to all reference techniques, but may be violated as a function of age and/or disease. In terms of equilibrium, one should be careful when selecting the time frame of interest, as this should coincide with the transient equilibrium of the tracer in all subjects.

The SUVR also depends on the rate of peripheral clearance of the tracer; unlike parameters derived from most kinetic models of brain uptake and binding, SUVR is not purely a function of brain parameters, though the extent to which differences in clearance between subjects affects study results has not been carefully examined for [^{11}C]DASB.

Studies that have used SUVR include Lee et al.,[65] Hesse et al.,[66] Ginovart et al.[5] and Houle et al.[3] To sum up, nine different methods have been applied to quantify [^{11}C]DASB PET. In our analysis, we find that the choice of kinetic model was associated with between-subject variability of ACC BP_{ND}. SRTM and non-invasive Logan (with Muller-Gartner PVC) produced the highest between-subject variabilities (Figure S4). When adding BP_{ND} as a covariate in the analysis, we also found a trend for a positive association ($P = 0.11$) between variation and BP_{ND}, highlighting a potential bias-variance trade-off in ACC BP_{ND} (Figure S5).

The identified bias-variance trade-off as a function of neuroimaging workflow warrants further investigation.

Test–retest studies for [^{11}C]DASB PET

To date, three test–retest studies for [^{11}C]DASB PET imaging have been published.[26,56,67] These studies involve two different scanners (ECAT HR + and GE Advance), one fixed time duration (120 min), two different dynamic framings (21 and 33 frames) and a range of 185 MBq to 740 MBq in injected dose. The studies included between 8 to 11 healthy subjects (aged 18–50) with a nearly 50/50 gender distribution. All test–retest scans were performed on the same day.

Two out the three studies used an AIF for the kinetic modeling, whereas one study used MRTM2 with cerebellum as reference region. Ogden et al.[26] reported that 100 min of scanning time was sufficient to obtain stable parameter estimates, and that the LEGA kinetic modeling approach produced the best results. However, the LEGA method produced a median percent difference in test–retest binding of approximately 20% ($n = 11$, range: 11–39.6%), when taken across all subjects and all VOIs. The median intraclass correlation coefficient (ICC) was approximately ICC = 0.8 (range: 0.455–0.926), taken across all subjects and all VOIs, with the highest ICC's in the dorsal caudate, thalamus and midbrain. Frankle et al.[56] obtained slightly higher ICCs compared to Ogden et al.[26] with a median ICC of 0.93 ($n = 9$, range: 0.79–0.97). Kim et al.[67] investigating the reproducibility of [^{11}C]DASB binding modeled with MRTM2 ($n = 8$), also used ICC as performance metric for test–retest reliability, including the additional performance metrics test–retest bias and test–retest variability. The results showed a significant negative bias in binding across test–retest, and high test–retest reliability for regions such as striatum, thalamus, temporal cortex and occipital cortex (ICC = 0.84). In contrast, poor test–retest reliability measures were obtained in the raphe and frontal cortex (ICC = 0.445). The reported negative bias across test–retest was barely discussed by Kim et al. and neither Frankle et al. nor Ogden et al. observed a negative test–retest bias with lower binding at retest.

The overall conclusion by Kim et al. was that the MRTM2 was reproducible and reliable for [^{11}C]DASB studies.

Notably, these test–retest studies were all performed on a relatively small sample and they demonstrate that some methods (i.e. Ogden et al. (LEGA) and Kim et al. (MRTM2)) are better or equally performing compared to other methods. However, the chosen performance metrics are not consistent across test–retest studies, and no attempt is explicitly made to address possible interactions with other preprocessing steps and/or other steps of the workflow (i.e. subject selection and data acquisition), as data acquisition and preprocessing are not consistent across the three test–retest studies. For example, Kim et al. used a summed image to perform frame-based MC, whereas Frankle et al. used a reference frame. However, while Frankle et al. used VOIs manually determined on MRIs according to well-defined operational criteria in conjunction with automated gray/white/CSF segmentation in cortex, Kim et al. instead used manual delineations without specifying the operational critera to obtain the VOIs. In addition, even though Frankle et al. and Ogden et al. used the same PET scanner to acquire the data, images were recontructed into 1.7 × 1.7 × 2.4 mm (non-isotropic) and 2.5 × 2.5 mm, respectively, with no specification on the z-direction in the latter study. All these modifications from study to study, make it difficult for the reader to infer whether reported methodological improvements are causally related to the new proposed method, or if it is due to a difference in data acquisition and/or preprocessing, limiting the generalizability to other neuroimaging workflows and studies.

Conclusions

In this review, we highlight the remarkable variety of ways in which researchers are currently performing complex neuroimaging studies, while implicitly expecting generalizable results across research groups. We systematically reviewed 105 published [^{11}C]DASB studies from 21 different PET centres, outlining differences in subject selection, data acquisition and preprocessing. Data sharing initiatives may significantly contribute to the understanding of the generalizable impact on such complex workflows, as the combined effects resulting from subject selection, data acquisition and preprocessing are unclear. We still need to understand the importance of bias-variance tradeoffs in neuroimaging experiments, and how neuroimaging workflows can be optimized for particular neuroscientific questions. The purpose of this study was not to identify a definitive PET preprocessing pipeline, but rather to establish workflow-dependent effects on binding and variation.

It is to be expected that the application of a new preprocessing pipeline will lead to different absolute binding measures, but the important question is whether the outcome of a study (i.e. difference between patients and controls) will remain.

In order to evaluate the extent to which any of the methodological factors described in this review matters, one needs to consider the given study aims.

For example, if an investigator wishes to compare age effects on the serotonin transporter in the striatum between two studies, it might be tempting to use both data sets, given that the methodology is internally consistent. However, while the researcher may not be able to combine the two sets of data, he/she may be able to use the two data sets seperately, assuming that the derived parameters while different, are scalable.

For future data sharing initiatives, it would be beneficial in a large and complete data set across a large number of subjects to assess which differences the various methodological variations can lead to, e.g. how much of a difference does variation in scanner resolution impact on, e.g., the striatum.

Our review focused on the radioligand [^{11}C]DASB, but the same considerations underlying the [^{11}C]DASB workflows could be made for any given PET or SPECT radiotracer. The aim of our paper is to highlight the need for transparency, reproducibility and to support future data sharing opportunities in the PET neuroimaging community. It is our hope that this work can also be used as a tool for future studies to evaluate the extent to which a given study deviates significantly from the current literature. From the current literature, it can be difficult to infer whether an observed change is physiological, or if it is driven by changes in subject selection and/or data acquisition and/or preprocessing. Data acquisition and preprocessing pipelines and their experimental interactions seem to be an overlooked aspect in modern PET neuroscience, and we believe that such testing is necessary in order to reliably provide new insights into human brain function.

Funding

MN was supported by the National Institutes of Health (Grant 5R21EB018964-02), the Lundbeck Foundation (Grant R90-A7722), and the Independent Research Fund Denmark (DFF-1331-00109 & DFF-4183-00627).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplementary material

Supplementary material for this paper can be found at the journal website: http://journals.sagepub.com/home/jcb

References

1. Wilson AA, Ginovart N, Schmidt ME, et al. Novel radiotracers for imaging the serotonin transporter by positron emission tomography: synthesis, radiosynthesis, in vitro, and ex vivo evaluation of [^{11}C]-labelled 2-(phenylthio)araalkylamines. *J Med Chem* 2000; 43: 3103–3110.
2. Rylands AJ, Hinz R, Jones M, et al. Pre- and postsynaptic serotonergic differences in males with extreme levels of impulsive aggression without callous unemotional traits: a PET study using ^{11}C-DASB and ^{11}C-MDL100907. *Biolog Psychiatry* 2012; 72: 1004–1011.
3. Houle S, Ginovart N, Hussey D, et al. Imaging the serotonin transporter with positron emission tomography: initial human studies with [^{11}C]DAPP and [^{11}C]DASB. *Eur J Nucl Med* 2000; 27: 1719–1722.
4. Holte S, Eriksson L and Dahlbom M. A preliminary evaluation of the Scanditronix PC1048-15B brain scanner. *Eur J Nucl Med* 1989; 15: 719–721.
5. Ginovart N, Wilson AA, Meyer JH, et al. Positron emission tomography quantification of [^{11}C]-DASB binding to the human serotonin transporter: modeling strategies. *J Cereb Blood Flow Metab* 2001; 21: 1342–1353.
6. Ogawa K, Tateno A, Arakawa R, et al. Occupancy of serotonin transporter by tramadol: a positron emission tomography study with [^{11}C]DASB. *Int J Neuropsychopharmacol* 2014; 17: 845–850.
7. Matsumoto K, Kitamura K, Mizuta T, et al. Performance characteristics of a new 3-dimensional continuousemission and spiral-transmission high sensitivity and high-resolution PET camera evaluated with the NEMA NU 2-2001 standard. *J Nucl Med* 2006; 47: 83–90.
8. Ouchi Y, Yoshikawa E, Futatsubashi M, et al. Altered brain serotonin transporter and associated glucose metabolism in Alzheimer disease. *J Nucl Med* 2009; 50: 1260–1266.
9. Watanabe M, Shimizu K, Omura T, et al. A new high-resolution PET scanner dedicated to brain research. *IEEE Trans Nucl Sci* 2002; 49: 634–639.
10. Fischl B and Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci U S A* 2000; 97: 11050–11055.
11. Takano A, Suzuki K, Kosaka J, et al. A dose-finding study of duloxetine based on serotonin transporter occupancy. *Psychopharmacology* 2006; 185: 395–399.
12. Matsumoto R, Ichise M, Ito H, et al. Reduced serotonin transporter binding in the insular cortex in patients with obsessive-compulsive disorder: a [^{11}C]DASB PET study. *Neuroimage* 2010; 49: 121–126.
13. Kim JS, Lee JS, Lee DS, et al. Performance evaluation of Siemens CTI ECAT EXACT 47 PET scanner using NEMA NU2-2001. *IEEE Nucl Sci Symp Conf Record* 2004; 5: 3118–3120.
14. Tyrer AE, Levitan RD, Houle S, et al. Serotonin transporter binding is reduced in seasonal affective disorder following light therapy. *Acta Psychiat Scand* 2016; 134: 410–419.
15. Kim E, Howes OD, Park JW, et al. Altered serotonin transporter binding potential in patients with obsessive-compulsive disorder under escitalopram treatment: [^{11}C]DASB PET study. *Psychol Med* 2016; 46: 357–366.
16. Adam L-E, Zaers J, Ostertag H, et al. Performance evaluation of the whole-body PET scanner ECAT EXACT HR/sup + / following the IEC standard. *IEEE Transac Nucl Sci* 1997; 44: 1172–1179.
17. Khohlmyer S and Stearns C. NEMA NU2-2001 performance results for the GE Advance PET system. *IEEE Record* 2002; 100: 890–894.
18. Wienhard K, et al. The ECAT HRRT: performance and first clinical application of the new high resolution research tomograph. *IEEE Trans Nucl Sci* 2002; 49: 104–110.
19. van Velden FH, Kloet RW, van Berckel BN, et al. HRRT versus HR + human brain PET studies: an interscanner test-retest study. *J Nucl Med* 2009; 50: 693–702.

20. Strecker K, Wegner F, Hesse S, et al. Preserved serotonin transporter binding in de novo Parkinson's disease: negative correlation with the dopamine transporter. *J Neurol* 2011; 258: 19–26.
21. Bhagwagar Z, Murthy N, Selvaraj S, et al. 5-HTT binding in recovered depressed patients and healthy volunteers: a positron emission tomography study with [^{11}C]DASB. *Am J Psychiatr* 2007; 164: 1858–1865.
22. Politis M, Wu K, Loane C, et al. Serotonergic mechanisms responsible for levodopa-induced dyskinesias in Parkinson's disease patients. *J Clin Invest* 2014; 124: 1–10.
23. Lee JY, Seo S, Lee JS, et al. Putaminal serotonergic innervation. *Neurology* 2015; 85: 853–860.
24. Kim JH, Son YD, Kim JH, et al. Serotonin trans- porter availability in thalamic subregions in schizophrenia: a study using 7.0-T MRI with [^{11}C]DASB high-resolution PET. *Psychiatry Res* 2015; 231: 50–57.
25. Lanzenberger R, Kranz GS, Haeusler D, et al. Prediction of SSRI treatment response in major depression based on serotonin transporter inter- play between median raphe nucleus and projection areas. *Neuroimage* 2012; 63: 874–881.
26. Ogden RT, Ojha A, Erlandsson K, et al. In vivo quantification of serotonin transporters using [^{11}C]DASB and positron emission tomography in humans: modeling considerations. *J Cereb Blood Flow Metab* 2007; 27: 205–217.
27. Frankle WG, Huang Y, Hwang DR, et al. Comparative evaluation of serotonin transporter radioligands ^{11}C-DASB and ^{11}C-McN 5652 in healthy humans. *J Nucl Med* 2004; 45: 682–694.
28. Morimoto T, Ito H, Takano A, et al. Effects of image reconstruction algorithm on neurotransmission PET studies in humans: comparison between filtered backprojection and ordered subsets expectation maximization. *Ann Nucl Med* 2006; 20: 237–243.
29. Bélanger MJ, Mann JJ and Parsey RV. OS-EM and FBP reconstructions at low count rates: effect on 3D PET studies of [^{11}C] WAY-100635. *Neuroimage* 2004; 21: 244–250.
30. Boellaard R, van Lingen A and Lammertsma AA. Experimental and clinical evaluation of iterative reconstruction (OSEM) in dynamic PET: quantitative characteristics and effects on kinetic modeling. *J Nucl Med* 2001; 42: 808–817.
31. James GM, Baldinger-Melich P, Philippe C, et al. Effects of selective serotonin reuptake inhibitors on interregional relation of serotonin transporter availability in major depression. *Front Hum Neurosci* 2017; 11: 48.
32. Frokjaer VG, Pinborg A, Holst KK, et al. Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: a positron emission tomography study. *Biol Psychiatr* 2015; 78: 534–543.
33. Cannon DM, et al. Elevated serotonin transporter binding in major depressive disorder assessed using positron emission tomography and [^{11}C]DASB: comparison with bipolar disorder. *Biol Psychiatr* 2007; 62: 870–877.
34. Vanicek T, Kutzelnigg A, Philippe C, et al. Altered interregional molecular associations of the serotonin transporter in attention deficit/hyperactivity disorder assessed with PET. *Hum Brain Mapp* 2017; 38: 792–802.
35. Politis M, Wu K, Loane C, et al. Serotonin neuron loss and nonmotor symptoms continue in Parkinson's patients treated with dopamine grafts. *Sci Transl Med* 2012; 4: 128ra41.
36. Greve DN, Salat DH, Bowen SL, et al. Different partial volume correction methods lead to different conclusions: an ^{18}F-FDG-PET study of aging. *Neuroimage* 2016; 132: 334–343.
37. Ganz M, Feng L, Hansen HD, et al. Cerebellar heterogeneity and its impact on PET data quantification of 5-HT receptor radioligands. *J Cereb Blood Flow Metab* 2017; 37: 3243–3252.
38. Beliveau V, Svarer C, Frokjaer VG, et al. Functional connectivity of the dorsal and median raphe nuclei at rest. *Neuroimage* 2015; 116: 187–195.
39. Frokjaer VG, Vinberg M, Erritzoe D, et al. High familial risk for mood disorder is associated with low dorsolateral prefrontal cortex serotonin transporter binding. *Neuroimage* 2009; 46: 360–366.
40. Savli M, Bauer A, Mitterhauser M, et al. Normative database of the serotonergic system in healthy subjects using multi-tracer PET. *Neuroimage* 2012; 63: 447–459.
41. Hinz R, Selvaraj S, Murthy NV, et al. Effects of citalopram infusion on the serotonin transporter binding of [^{11}C]DASB in healthy controls. *J Cereb Blood Flow Metab* 2008; 28: 1478–1490.
42. Takano A, Suzuki K, Kosaka J, et al. A dose-finding study of duloxetine based on serotonin transporter occupancy. *Psychopharmacology* 2006; 185: 395–399.
43. Zientek F, Winter K, Müller A, et al. Effortful control as a dimension of temperament is negatively associated with prefrontal serotonin transporter availability in obese and non-obese individuals. *Eur J Neurosci* 2016; 44: 2460–2424.
44. Roussakis AA, Politis M, Towey D, et al. Serotonin-to-dopamine transporter ratios in Parkinson disease. *Neurology* 2016; 86: 1152–1158.
45. Poldrack RA, Fletcher PC, Henson RN, et al. Guidelines for reporting an fMRI study. *Neuroimage* 2008; 40: 409–414.
46. Frokjaer VG, Erritzoe D, Holst KK, et al. Prefrontal serotonin transporter availability is positively associated with the cortisol awakening response. *Eur Neuropsychopharmacol* 2013; 23: 285–294.
47. Frokjaer VG, Erritzoe D, Holst KK, et al. In abstinent MDMA users the cortisol awakening response is off-set but associated with prefrontal serotonin transporter binding as in non-users. *Int J Neuropsychopharmacol* 2014; 17: 1119–1128.
48. Marner L, Frokjaer VG, Kalbitzer J, et al. Loss of serotonin 2A receptors exceeds loss of serotonergic projections in early Alzheimer's disease: a combined [^{11}C]DASB and [^{18}F]altanserin-PET study. *Neurobiol Aging* 2012; 33: 479–487.
49. Erlandsson K, Buvat I, Pretorius PH, et al. A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. *Phys Med Biol* 2012; 57: R119–R159.

50. Nogami T, Takano H, Arakawa R, et al. Occupancy of serotonin and norepinephrine transporter by milnacipran in patients with major depressive disorder: a positron emission tomography study with [^{11}C]DASB and (S,S)-[^{18}F]FMeNER-D(2). *Int J Neuropsychopharmacol* 2013; 16: 937–943.
51. Miller JM, Everett BA, Oquendo MA, et al. Positron emission tomography quantification of serotonin transporter binding in medication-free bipolar disorder. *Synapse* 2016; 70: 24–32.
52. Parsey RV, Kent JM, Oquendo MA, et al. Acute occupancy of brain serotonin transporter by sertraline as measured by [^{11}C]DASB and positron emission tomography. *Biol Psychiatr* 2006; 59: 821–828.
53. Shapiro PA, Sloan RP, Deochand C, et al. Quantifying serotonin transporters by PET with [^{11}C]-DASB before and after interferon-alpha treatment. *Synapse* 2014; 68: 548–555.
54. Kish SJ, Furukawa Y, Chang LJ, et al. Regional distribution of serotonin transporter protein in postmortem human brain: is the cerebellum a SERT-free brain region? *Nucl Med Biol* 2005; 32: 123–128.
55. McCann UD, Szabo Z, Seckin E, et al. Quantitative PET studies of the serotonin trans- porter in MDMA users and controls using [^{11}C]McN5652 and [^{11}C]DASB. *Neuropsychopharmacology* 2005; 30: 1741–1750.
56. Frankle WG, Slifstein M, Gunn RN, et al. Estimation of serotonin transporter parameters with ^{11}C-DASB in healthy humans: reproducibility and comparison of methods. *J Nucl Med* 2006; 47: 815–826.
57. van de Giessen E, Rosell DR, Thompson JL, et al. Serotonin transporter availability in impulsive aggressive personality disordered patients: a PET study with [^{11}C]DASB. *J Psychiatr Res* 2014; 58: 147–154.
58. Murthy NV, Selvaraj S, Cowen, et al. Serotonin transporter polymorphisms (SLC6A4 insertion/deletion and rs25531) do not affect the availability of 5-HTT to [^{11}C]DASB binding in the living human brain. *Neuroimage* 2010; 52: 50–54.
59. 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. *J Cereb Blood Flow Metab* 2003; 23: 1096–1112.
60. Fisher PM, Ozenne B, Svarer C, et al. BDNF val66met association with serotonin transporter binding in healthy humans. *Transl Psychiatr* 2017; 7: e1029.
61. Lammertsma AA and Hume SP. Simplified reference tissue model for PET receptor studies. *Neuroimage* 1996; 4: 153–158.
62. Wu Y and Carson RE. Noise reduction in the simplified reference tissue model for neuroreceptor functional imaging. *J Cereb Blood Flow Metab* 2002; 22: 1440–1452.
63. 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: 834–840.
64. Yoder KK, Territo PR, Hutchins GD, et al. Comparison of standardized uptake values with volume of distribution for quantitation of [^{11}C]PBR28 brain uptake. *Nucl Med Biol* 2015; 42: 305–308.
65. Lee JY, Seo S, Lee JS, et al. Putaminal serotonergic innervation. *Neurology* 2015; 85: 853–860.
66. Hesse S, Brust P, Mäding P, et al. Imaging of the brain sero- tonin transporters (SERT) with ^{18}F-labelled fluoromethyl- McN5652 and PET in humans. *Eur J Nucl Med Mol Imag* 2012; 39: 1001–1011.
67. Kim JS, 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: 208–214.
68. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013; 14: 365–376.
69. Moses-Kolko EL, Price JC, Shah N, et al. Age, sex, and reproductive hormone effects on brain serotonin-1A and serotonin-2A receptor binding in a healthy population. *Neuropsychopharmacology* 2011; 36: 2729–2740.
70. Jovanovic H, Lundberg J, Karlsson P, et al. Sex differences in the serotonin 1A receptor and serotonin transporter binding in the human brain measured by PET. *Neuroimage* 2008; 39: 1408–1419.
71. Liu X, Cannon DM, Akula N, et al. A non-synonymous polymorphism in galactose mutarotase (GALM) is associated with serotonin transporter binding potential in the human thalamus: results of a genome-wide association study. *Mol Psychiatr* 2011; 16: 584–594.
72. Tuominen L, Miettunen J, Cannon DM, et al. Neuroticism associates with cerebral in vivo serotonin transporter binding differently in males and females. *Int J Neuropsychopharmacol* 2017; 20: 963–970.
73. van den Heuvel OA, Boellaard R, Veltman DJ, et al. Attenuation correction of PET activation studies in the presence of task-related motion. *Neuroimage* 2003; 19: 1501–1509.
74. Boileau I, Warsh JJ, Guttman M, et al. Elevated serotonin transporter binding in depressed patients with Parkinson's disease: a preliminary PET study with [^{11}C]DASB. *Mov Disord* 2008; 23: 1776–1780.
75. Carson RE, Barker C, Liow JS, et al. Design of a motion-compensation OSEM List-mode Algorithm for resolution-recovery Reconstruction for the HRRT. *IEEE Nucl Sci Symp Conf Record* 2003; 5: 3281–3285.
76. Olesen OV, Sullivan JM, Mulnix T, et al. List-mode PET motion correction using markerless head tracking: proof-of-concept with scans of human subject. *IEEE Trans Med Imag* 2013; 32: 200–209.

Appendix B

Paper [B]

Nørgaard M, Ganz M, Svarer C, Vibe G. Frokjaer, Greve DN, Strother SC, Knudsen GM. Optimization of Preprocessing Strategies in Positron Emission Tomography (PET) Neuroimaging: A [^{11}C]DASB Study. In revision, *NeuroImage*.

Optimization of Preprocessing Strategies in Positron Emission Tomography (PET) Neuroimaging: A [^{11}C]DASB PET Study

Running Title: Preprocessing Effects in PET Neuroimaging

Martin Nørgaard[1,2]
Melanie Ganz[1,3]
Claus Svarer[1]
Vibe G. Frokjaer[1]
Douglas N. Greve[5]
Stephen C. Strother[4]
Gitte M. Knudsen[1,2*]

[1] Neurobiology Research Unit, Copenhagen University Hospital
Rigshospitalet, Copenhagen, Denmark
[2] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[3] Faculty of Computer Science, University of Copenhagen, Copenhagen, Denmark
[4] Rotman Research Institute at Baycrest, and Department of Medical Biophysics, University of Toronto, Toronto, Canada
[5] Athinoula A. Martinos Center for Biomedical Imaging,
Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

* Corresponding author gmk@nru.dk

Abstract

Positron Emission Tomography (PET) is an important neuroimaging tool to quantify the distribution of specific molecules in the brain. The quantification is based on a series of individually designed data preprocessing steps (pipeline) and an optimal preprocessing strategy is per definition associated with less noise and improved statistical power, potentially allowing for more valid neurobiological interpretations. In spite of this, it is currently unclear how to design the best preprocessing pipeline and to what extent the choice of each preprocessing step in the pipeline minimizes subject-specific errors.

To evaluate the impact of various preprocessing strategies, we examined 384 different pipeline strategies in data from 30 healthy participants scanned twice with the serotonin transporter (5-HTT) radioligand [^{11}C]DASB. Five commonly used preprocessing steps with two to four options were investigated: (1) motion correction (MC) (2) co-registration (3) delineation of volumes of interest (VOI's) (4) partial volume correction (PVC), and (5) kinetic modeling. To quantitatively compare and evaluate the impact of various preprocessing strategies, we used the performance metrics: test-retest bias, within- and between-subject variability, the intraclass-correlation coefficient, and global signal-to-noise ratio. We also performed a power analysis to estimate the required sample size to detect either a 5% or 10% difference in 5-HTT binding as a function of preprocessing pipeline.

The results showed a complex downstream dependency between the various preprocessing steps on the performance metrics. The choice of MC, PVC, and kinetic modeling had the most profound effects on 5-HTT binding, and the effects differed across VOI's. Notably, we observed a negative bias in 5-HTT binding across test and retest in 98% of pipelines, ranging from 0-6% depending on the pipeline. Optimization of the performance metrics revealed a trade-off in within- and between-subject variability at the group-level with opposite effects (i.e. minimization of within-subject variability increased between-subject variability and vice versa). The sample size required to detect a given effect size was also compromised by the preprocessing strategy, resulting in up to 80% increases in sample size needed to detect a 5% difference in 5-HTT binding.

This is the first study to systematically investigate and demonstrate the effect of choosing different preprocessing strategies on the outcome of dynamic PET studies. We show how optimal and maximally powered neuroimaging results can be obtained by choosing appropriate preprocessing strategies and we provide recommendations depending on the study design.

In addition, the results contribute to a better understanding of methodological uncertainty and variability in preprocessing decisions for future group- and/or longitudinal PET studies.

Key words: Positron Emission Tomography; preprocessing; head motion; optimization; partial volume correction; kinetic modeling; test-retest; [^{11}C]DASB

INTRODUCTION

Positron Emission Tomography (PET) is a state-of-the-art neuroimaging tool for quantification of the *in vivo* spatial distribution of specific molecules in the brain. It has long been recognized that precise quantification using a series of individually designed preprocessing steps ("pipeline") is a critical part of a PET analysis framework, and as part of the validation of new PET radioligands, these are often preprocessed with different kinetic models and at different scan lengths. The outcomes are often examined in a test-retest setting (Parsey et al. 2000, Ginovart et al. 2001) under the implicit assumption that test and retest should generate similar outcomes. However, despite the importance and usefulness of validating kinetic models and scan length, the impact of several other important choices 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), remain unresolved.
As a result, centres or even individual scientists often design their own unique preprocessing strategy (Nørgaard et al. 2018), with each choice potentially compromising one another to affect performance (e.g. NRM Grand Challenge 2018, www.petgrandchallenge.com).

The first preprocessing step of a common PET preprocessing workflow is often motion correction (MC), which intends to remove head motion artefacts from the PET data originating from the data acquisition. However, there is currently no consensus in the literature about the value of adding MC. For example, 40% of published [^{11}C]DASB studies abstain from conducting MC, and the remaining studies apply it in different forms (Nørgaard et al. 2018). Even if head motion is minimal, the extent to which this step affects the spatiotemporal signal distribution is largely unknown, mainly because only a few studies have investigated its impact (e.g., Montgomery et al. 2006).

The second preprocessing step, co-registration, integrates the anatomical information from the MRI with the functional PET. Unless a precise co-registration is obtained, the two modalities will not overlap correctly (Schwarz et al. 2017). Several well-established techniques are available for the purpose, such as Normalized Mutual Information (NMI; Studholme et al. 1999) or Boundary-Based Registration (BBR; Greve et al. 2009), but there is currently no consensus as to which PET image should be chosen for the anatomical alignment. Some groups use the time-weighted average PET image (weighting each frame with the SNR, e.g. Frøkjaer et al. 2015), whereas others use the average PET image (weighting each frame equally, e.g. Frick et al. 2015).

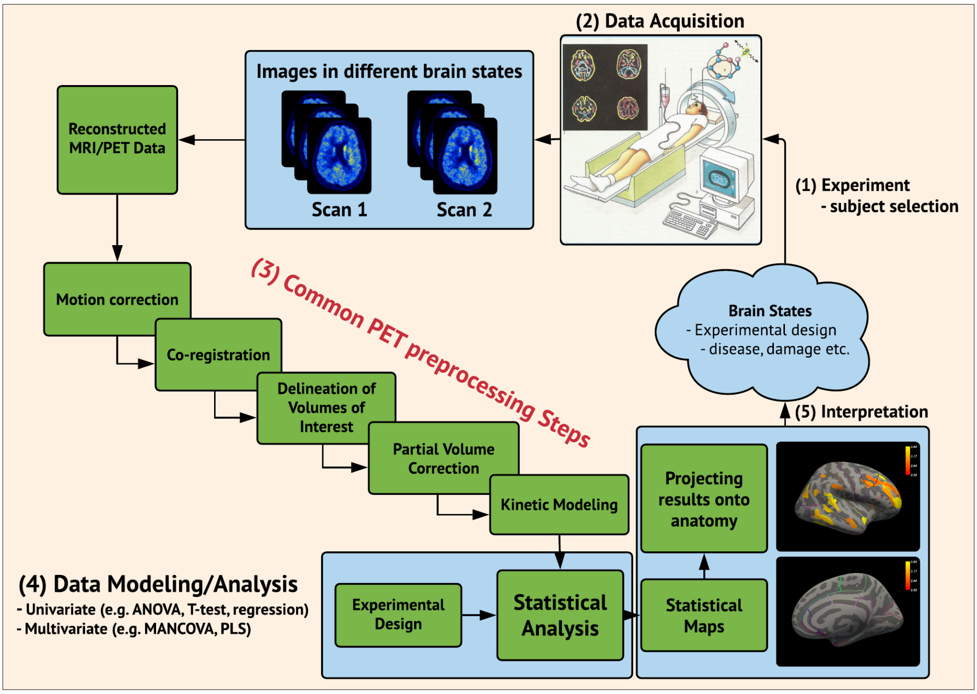
(2) Data Acquisition
Images in different brain states
Scan 1
Scan 2
Reconstructed MRI/PET Data
(1) Experiment
- subject selection
Motion correction
(3) Common PET preprocessing Steps
Co-registration
Delineation of Volumes of Interest
Partial Volume Correction
Kinetic Modeling
Brain States
- Experimental design
- disease, damage etc.
(5) Interpretation
Projecting results onto anatomy
(4) Data Modeling/Analysis
- Univariate (e.g. ANOVA, T-test, regression)
- Multivariate (e.g. MANCOVA, PLS)
Experimental Design
Statistical Analysis
Statistical Maps

Figure 1: Flowchart depicting a common pipeline for neuroimaging studies (multimodal PET and MRI) and its multiple stages ranging from (1) experimental design / subject selection, (2) data acquisition, (3) preprocessing, (4) data modeling/analysis, and (5) interpretation.

The third preprocessing step, delineation of VOI's, is most often the next logical preprocessing step to be carried out, as hypotheses related to brain function often are region specific. Such delineations can either be drawn manually on an MRI (Roussakis et al. 2016) or PET (Meyer et al. 2001), or by using automatic software packages such as FreeSurfer (Fishl et al. 2012; https://surfer.nmr.mgh.harvard.edu/), SPM (http://www.fil.ion.ucl.ac.uk/spm/) or PVElab (Svarer et al. 2005). However, when delineating VOI's manually not all studies use well-defined operational criteria and are blind to subject diagnosis. In addition, VOI's are often delineated under the assumption of homogeneously distributed radioligand within the target volume, and if this assumption is violated it can potentially affect the outcome measure (Schain et al. 2014, Nørgaard et al. 2015).

Despite its intention to correct for partial volume effects originating from limited spatial resolution, the fourth preprocessing step, PVC, is only rarely used in PET studies. For example, only 4 out of 105 published [^{11}C]DASB studies used PVC (Nørgaard et al. 2018) with no validation studies (i.e., test-retest) currently published for [^{11}C]DASB. While PVC can cause noise amplification (Greve et al. 2014), it must be used in studies where there are expected differences or changes in brain atrophy as these can cause a false change in the PET signal (Greve et al. 2016). Greve et al. 2016 suggested the Geometric Transfer Matrix (GTM) to be the preferred method for VOI analysis compared to no PVC, but this remains to be validated in a test-retest setting.

For quantification of dynamic PET images (fifth preprocessing step), different kinetic models are commonly used to provide information about the binding, e.g., the non-displaceable binding potential, $BP_{ND,}$ in different brain regions (Ichise et al. 2003, Ginovart et al. 2001, Innis et al. 2007). Whereas the pros and cons of reference tissue based methods (Ganz et al. 2017, Frick et al. 2015) versus arterial input methods (Parsey et al. 2006, Rylands et al. 2012) are up for discussion, the reference tissue modeling (RTM) approach comes with some obvious benefits for the patient (no arterial line) and for the staff (no blood measurements of radioactivity and metabolites). The gold standard is to acquire arterial blood samples in combination with the PET scan, providing an arterial input function for subsequent kinetic modeling. However, while an arterial input function ideally generates an unbiased estimate of the radioligand binding, noisy estimates of count rate and correction of radiometabolites generally add additional variation into the data. Hence, once a radioligand has been favorably validated with arterial input function against RTM, the latter is often used instead, e.g., the Simplified Reference Tissue Model (SRTM) by Lammertsma and Hume 1996, the extended version SRTM2 by Wu and Carson 2003, the non-invasive Logan procedure by Logan et al. 1996, the Multilinear-Reference Tissue Model (MRTM) by Ichise et al. 2003, and the extended version MRTM2 by Ichise et al. 2003. These RTM's mainly differ in the model-parameter estimation, and how the noise is controlled. The kinetic modeling step has been thoroughly investigated in the PET literature, showing different performances across VOIs and subjects (Ginovart et al. 2001, Ogden et al. 2006, Kim et al. 2006).

In this study, we extend the work of previous validation studies of the radioligand [^{11}C]DASB, which binds to the serotonin transporter (5-HTT), a target for many anti-depressive drugs (Houle et al. 2000, Meyer et al. 2001). Previously published [^{11}C]DASB PET papers have mainly used five preprocessing steps with multiple levels of options within each preprocessing step (Figure 1). However, while there is some consensus on the main preprocessing steps (MC, co-registration, VOI, PVC, and kinetic modeling), there is less consensus on the details within each step. In addition, with new methodological improvements continually being developed and refined (Zanderigo et al. 2017, Gryglewski et al. 2017) it may also be difficult to establish an optimal pipeline, with each choice potentially compromising one another. Nevertheless, for scientific, ethical and economical reasons it is important to know how the choice of preprocessing strategy influences the noise levels and thereby the sample size required to establish e.g. group differences. Inspired by the previously published preprocessing strategies for the radioligand [^{11}C]DASB (Nørgaard et al. 2018), in this work we will focus on three key questions:

(1) are measures of 5-HTT BP_{ND} using [^{11}C]DASB robustly determined across a wide range of preprocessing strategies? The robustness will be estimated using the performance metrics; test-retest bias, within- and between-subject variability, global signal-to-noise ratio (*gSNR*), and intraclass correlation coefficient (Kim et al. 2006, Strother et al. 2002).

(2) does optimization of the performance metrics result in a detectable tradeoff in within- and between-subject variability at the group level?

(3) can study power be enhanced by optimized preprocessing of [^{11}C]DASB?

We specifically chose to focus on the PET radioligand [^{11}C]DASB because of its widespread use to study various aspects of brain function, but more importantly because the foundation for selecting a given preprocessing strategy seems to be an overlooked aspect in modern PET neuroscience.

MATERIALS AND METHODS

1.1 Participants

A total of $N=30$ healthy women (mean age: 25 ± 5.9 years, range: 18 – 37) were recruited 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 treatment only (a saline injection), i.e., the data is considered to represent test-retest without any expected changes in [^{11}C]DASB binding. The study by Frokjaer et al. 2015 was designed to capture brain chemistry in two consecutive follicular phases of the menstrual cycle and participants were therefore planned to be re-scanned 23-35 days after their baseline cycle scan (depending on their follow-up cycle). Three participants were scanned one cycle-period later (61 days, 70 days, 56 days), one participant two periods later (92 days), and one participant three periods later (122 days). The midfollicular timing of the scan was kept in all participants. All the remaining 25 participants were scanned in a cycle-period ranging between 27 and 37 days. In addition, participants were scanned at similar time of the day in scan 1 and scan 2, eliminating potential diurnal effects. 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.

1.2 Magnetic Resonance Imaging Acquisition

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° was acquired for all patients using a Siemens Magnetom Trio 3T MR scanner or a Siemens 3T Verio MR scanner. In addition, a 3D T2-weighted isotropic sagittal sequence with matrix size 256 x 256 x 176; voxel size = 1 x 1 x 1 mm; TR/TE = 3200/409 ms; flip angle = 120˚ was also acquired for all subjects. All single-subject MRI sequences were corrected for gradient nonlinearities according to Jovicich et al. 2006, in order to correct for spatial distortions and achieve optimal PET-MR co-registration. All the acquired MR images were examined for structural abnormalities, as a criterion for subject inclusion.

1.3 Positron Emission Tomography using [^{11}C]DASB

All patients 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 ± 30 (mean ± 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).

1.4 Preprocessing Steps for PET and MRI

Here we establish a 5-step pipeline, each step with 2 to 4 options, to estimate the outcome measure BP_{ND}. All the individual procedures have previously been used in published [^{11}C]DASB PET studies, except for PVC using the GTM. The steps are listed below in the order in which they were applied. Specific rationales for including/excluding each unique preprocessing step and their options are listed below.

Step 1 – Motion correction (2 options)

Within-scan PET motion correction was executed using a data-driven automated image registration (AIR v. 5.2.5, http://loni.usc.edu/Software/AIR). Prior to alignment, each frame was smoothed using a 10 mm Gaussian 3D kernel and thresholded at the 20-percentile level to boost SNR. Alignment parameters were estimated for the smoothed PET frames 10-36 to a reference frame with high SNR (frame 26) using a scaled least squares cost-function in AIR. Subsequently, the non-smoothed frames were transformed using the estimated alignment parameters and resliced into a 4D motion corrected data set (e.g., as applied in Frokjaer et al. 2015 and Beliveau et al. 2017). The motion correction estimation for frame 10 was applied to the first 9 frames. We chose to register frames 10-36 only, because the first 9 time frames (10/20 sec) have low count statistics, high noise levels and have shown to produce highly variable alignment parameters.
Criterion for acceptable motion was a median movement less than 3 mm across frames, as estimated by the median of the sum of the squared translations (x,y,z) across all voxels.
All 30 participants had acceptable median motion below 3 mm.
The rationale for testing the effect of MC in the pipeline is because motion artefacts vary by dataset. Furthermore, MC should ultimately control motion artefacts, but may also impose unwanted biases on the data or reduce experimental power, especially in cases of minor or no head movement (Churchill et al. 2012). In addition, Nørgaard et al. 2018 showed that MC lowers between subject variability in striatum, resulting in 26% fewer subjects needed in a group analysis to achieve similarly power statistical tests. It is therefore of interest to validate this observation in an independent dataset.

Step 2 – Co-registration (4 options)

All single-subject PET frames were initially either summed (according to their frame length i.e. integral) 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 step is explicitly evaluated, as its effects may vary by dataset and as a function of SNR.

Step 3 – Delineation of Volumes of Interest (3 options)

All MRI scans were processed using FreeSurfer (http://surfer.nmr.mgh.harvard.edu, version 5.3). FreeSurfer contains a fully automatic structural imaging pipeline for processing of cross-sectional as well as longitudinal data. Furthermore, it includes several features such as skull stripping, B1 bias field correction, non-linear registration to a stereotaxic atlas, statistical analysis of morphometric differences, and probabilistic labeling of cortical/subcortical brain structures based on the Desikan-Killiany atlas (Fischl et al. 2004). A total of 28 subcortical and cortical regions were extracted, and averaged across hemispheres producing a final sample of 14 regions pr. subject/pipeline. The volumetric regions included the *amygdala, thalamus, putamen, caudate, anterior cingulate cortex (ACC), hippocampus, orbital frontal cortex, superior frontal cortex, occipital cortex, superior temporal gyrus, insula, inferior temporal gyrus, parietal cortex, and entorhinal cortex*. We chose these regions because they largely cover the entire brain, but also because many of the regions have been used in previously published DASB PET studies. Out of more than 100 published DASB PET studies (Nørgaard et al. 2018), each region is mentioned N times: amygdala (N=72), thalamus (N=105), putamen (N=88), caudate (N=82), ACC (N=74), hippocampus (N=71), frontal cortex (N=66), occipital cortex (N=48), temporal cortex (N=58), parietal cortex (N=34), entorhinal cortex (N=16). Subsequently to running the FreeSurfer pipeline, the researcher can choose to perform user-dependent *manual edits* to the FreeSurfer output, to correct for errors mostly located in the white matter (WM), cerebrospinal fluid (CSF) or on the pial surface. The manual editing was carried out according to FreeSurfer recommendations (https://surfer.nmr.mgh.harvard.edu/fswiki/Edits).

If a T2-weighted MRI is also available, semi user-independent edits can also be made to the FreeSurfer output by re-running the FreeSurfer reconstruction including the T2-weighted MRI. We examined all three pipelines in this study and now refer to these as **FS-RAW** (standard output from FreeSurfer), **FS-MAN** (output from FreeSurfer with manual edits) and **FS-T2P** (output from FreeSurfer with the T2 stream). Only the first test-scan MRI was used for the analysis. Different FreeSurfer options are tested, as the optimal correction has been reported to vary as a function of subject and scanner (McCarthy et al. 2015). Although choice of atlas (e.g. PVElab, AAL or

MNI305) may have an impact on the outcome, we considered assessment of various atlas choices to be beyond the scope of the current work and we consistently applied the Desikan-Killiany atlas provided in FreeSurfer. However, as it is also common to include a normalization step to standard space and subsequently extract VOIs using a volumetric atlas, an evaluation and comparison of such a pipeline can be found in the supplemental material.

Step 4 – Partial Volume Correction (4 options)

The data were analyzed either without or with three partial volume correction (PVC) approaches. The VOI-based PVC technique, Geometric Transfer Matrix (GTM), by Rousset et al. 1998 was applied using PETsurfer (https://surfer.nmr.mgh.harvard.edu/fswiki/PetSurfer), 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 reconstructed with a OP-OSEM-PSF algorithm varies from 1-2.5 mm in radial orientation depending on the distance from the centre of the field of view (Olesen et al. 2009), we ran the analyses with the PSF settings; 0 mm and 2 mm. However, because motion, inhomogeneous tracer uptake and varying uptake across frames is likely to further decrease the spatial resolution as compared to a point source in Olesen et al. 2009, we also ran the PVC analyses with 4 mm, as used in Greve et al. 2014. The PVC step is evaluated, because it has been suggested to be the optimal solution for VOI analysis, given that assumptions about the PSF, accurate delineation of regions, correct PET-MRI registration, and constant uptake within each VOI are satisfied (Greve et al. 2016). In addition, a homogeneous CSF and WM segmentation is important (provided in FreeSurfer), as these are primary regions to compensate for gray matter uptake of the tracer. When the assumptions are satisfied (and under noiseless conditions), the GTM will provide the exact mean in each VOI.

Step 5 – Kinetic Modeling (4 options)

The Multilinear Reference Tissue Model (MRTM) was applied as described by Ichise et al. 2003 with cerebellum (excluding vermis) as a reference region, allowing for estimation of three parameters from which the non-displaceable binding potential (BP_{ND}) can be derived.

The second model applied was the Multilinear Reference Tissue Model 2 (MRTM2) (Ichise et al. 2003) with cerebellum (excluding vermis) as a reference region, and thalamus, putamen and caudate were averaged to represent a single less noisy high-binding region for estimation of k_2', the clearance rate constant from reference region to plasma (Beliveau et al. 2016). The MRTM2 is similar to MRTM, except that k_2' is determined after the first iteration of MRTM and its value is subsequently entered into the two-parameter MRTM2 model. This approximates a linear kinetic analysis, but is executed in only a fraction of the computational time. The simplified reference tissue model, SRTM, was applied as described by Lammertsma and Hume, 1996. SRTM allows for nonlinear least squares estimation of 3 parameters (R_1, k_2' and k_{2a}) from the full dataset, and the $\mathrm{BP}_{\mathrm{ND}}$ can be estimated from $BP_{ND} = R_1\left(\frac{k_2'}{k_{2a}}\right) - 1$. R_1 is the relative radioligand delivery and k_{2a} is the apparent rate constant.

The non-invasive Logan reference tissue model was applied as described in Logan et al. 1996 with $t^* = 35$ minutes for all regions and subjects. It also assumes the existence of a valid reference region and an average tissue-to-plasma clearance k_2', and the distribution volume ratio can estimated as $DVR = BP_{ND} - 1$. All kinetic models applied in this work were implemented in MATLAB v. 2016b as specified in their original paper. The implementation in MATLAB was validated with PMOD v. 3.0 (10 subjects < 0.1% difference in $\mathrm{BP}_{\mathrm{ND}}$), but was carried out in MATLAB for parallel execution purposes to substantially reduce processing time.

Different kinetic modeling approaches are tested in this study, as the optimal estimation of 5-HTT binding may vary as a function of scanner (i.e. resolution), subject and region.

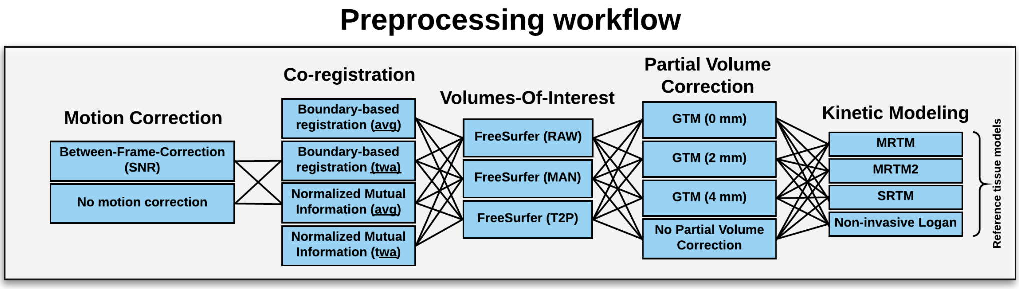

Figure 2: Schematic overview of the various preprocessing steps applied for the [^{11}C]DASB quantification. Abbreviations; average (avg), time-weighted average (twa), signal-to-noise ratio (SNR).

From this 5-step list of preprocessing choices, we can quantify $\mathrm{BP}_{\mathrm{ND}}$ for 3x2x4x4x4 = 384 different pipelines per subject (Figure 2) and subsequently examine their impact on a set of chosen performance metrics (Section 1.5).

1.5 Analysis and pipeline performance metrics

To evaluate the effects of different PET preprocessing choices, we tested their performance on a set of common performance metrics, namely the test-retest bias, within-subject variability, between-subject variability, intraclass correlation, power calculation, and failure rate. While these analyses were applied for each region *j* individually and summarized over all subjects *i*, we also adopted a global reproducibility metric from the fMRI literature, producing a single reproducibility measure for each subject *i* and pipeline *k*, taking the information from all regions into account (Strother et al. 2002). This sums to a total of 7 performance metrics that serve to assess the individual pipelines against each other.

Unlesss otherwise stated, we used statistical subsampling to test several sample sizes of either $\tilde{n}$ = 10 or 20 subjects randomly selected without replacement from the 30 subjects, and this was repeated 1000 times, to produce a mean estimate and a 95% confidence interval (CI). The sample sizes of 10 or 20 subjects were chosen to reflect the commonly used sample sizes in [^{11}C]DASB PET studies (Nørgaard et al. 2018). Notation-wise, $\tilde{n}$ indicates a resampling analysis, whereas N = 30 indicates that all subjects were included in the analysis.

Statistical differences in pipeline choice (e.g., motion correction vs. no motion correction) for each performance metric was determined across 1000 resamples (subsampling 20 subjects without replacement), and then using the empirical distribution of the differences of the performance metric. This provides an empirical P-value for the difference between pipeline choices for each performance metric. Correction for multiple comparisons across regions was carried out using False-Discovery Rate (FDR), at FDR=0.05. The rationale for choosing these 7 performance metrics is to provide a quantitative estimate of what can be expected of biases and variability as a function of preprocessing pipeline choice and sample size.

Global Within-Subject Reproducibility Metric (FIX)

A global within-subject reproducibility metric over all regions was estimated by generating global signal-to-noise (gSNR) metrics for each subject *i* and pipeline *k*, as described in (Strother et al. 2002, Churchill et al. 2012, Churchill et al. 2015). The fourteen brain regions, described in section 1.4 step 3, were selected for analysis, and a pairwise linear correlation based on the Pearson linear correlation coefficient, *R*, was estimated based on the test and retest BP_{ND}'s.

The *gSNR* for each subject and pipeline was estimated as

$$\mathrm{gSNR}_{i,k} = \sqrt{\frac{(1+\mathrm{R}_{i,k}) - (1-\mathrm{R}_{i,k})}{(1-\mathrm{R}_{i,k})}}$$

Subsequently, we identified the pipeline that maximized the median-rank across all subjects, as described in (Churchill et al. 2012), and described in Supplemental Text 1. This pipeline is defined as the optimal fixed pipeline (FIX) across all subjects and regions.

Test-retest Bias

The test-retest bias was estimated as the difference between the two measurements for subject *i*, region *j*, and pipeline *k,* and expressed as a percentage of the first measurement (Kim et al. 2006). This is given by

$$Bias_{i,j,k} = 100\times\left(\frac{retest_{i,j,k} - test_{i,j,k}}{test_{i,j,k}}\right)$$

In the estimation of an average group-level bias (i.e. $Bias_{j,k} = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} Bias_{i,j,k}$), $\mathrm{BP_{ND}}$'s that were ≤ 0 or ≥ 10 in either test or retest were excluded in the estimation to avoid the influence of outliers. To account for this exclusion, we estimated a "failure rate" (performance metric 7) for a given region *j* and pipeline *k,* defined as the number of outliers divided by the number of subjects x 100.

Within-Subject Variability Metric (WSV)

The within-subject variability (WSV) was estimated as the standard deviation across regions of the difference between test and retest. To normalize the metric to a coefficient of variation (CV) %, we divided the WSV by the average $\mathrm{BP_{ND}}$'s over test and retest for all 30 subjects (outliers excluded). This can mathematically be expressed as

$$CV_{j,k} = 100\times\left(\frac{\sqrt{\frac{\sum_{i=1}^{\tilde{n}}\left(d_{i,j,k} - \bar{d}_{j,k}\right)^2}{n-1}}}{\frac{\sum_{i=1}^{S}\left(test_{i,j,k} + retest_{i,j,k}\right)/2}{S}}\right)$$

where $d_{i,j,k} = test_{i,j,k} - retest_{i,j,k}$, $\bar{d}_{j,k} = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} d_{i,j,k}$ and S is the number of subsamples (i.e. without outliers). BP_{ND}'s that were ≤ 0 or ≥ 10 in either test or retest were excluded in the estimation to avoid the influence of outliers.

Between-Subject Variability Metric (BSV)

Between-subject variability (BSV) was captured by identifying the pipeline that minimized the mean standard deviation across all regions and across subjects at baseline (i.e. test). To compare regions, we estimated the CV by dividing the standard deviation, σ, for $\tilde{n}$ = 10 or 20 subjects for a given region by the mean, μ, estimated from all subjects at baseline and re-scan (outliers excluded). This is our final between-subject variability measure.

$$CV_{j,k} = 100\times\left(\frac{\sigma_{j,k}}{\mu_{j,k}}\right)$$

where $\mu_{j,k} = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(test_{i,j,k} + retest_{i,j,k})/2$ and $\sigma_{j,k} = \sqrt{\frac{1}{\tilde{n}-1}\sum_{i=1}^{\tilde{n}}((\frac{test_{i,j,k}+retest_{i,j,k}}{2}) - \mu_{j,k})^2}$.

BP_{ND}'s that were ≤ 0 or ≥ 10 in either test or retest were excluded in the estimation to avoid the influence of outliers.

Intraclass-Correlation Reproducibility Metric (ICC)

We estimated the intraclass-correlation coefficient (ICC(3,1)) across test-retest for all regions and for each pipeline, as given below

$$ICC_{j,k} = \frac{MSBS_{j,k} - MSE_{j,k}}{MSBS_{j,k} + (q-1)MSE_{j,k}}$$

where q is the number of repeated measurements (i.e. $q = 2$), MSBS is the between-subjects' sum of squares, and MSE is the error mean square. We chose ICC(3,1), as this measure eliminates possible systematic test-retest effects due to the scan-order, by treating repeated measurements as fixed instead of random. In the estimation of the ICC metric, BP_{ND}'s that were ≤ 0 or ≥ 10 in either test or retest were excluded to avoid the influence of outliers. We subsequently chose the pipeline, that maximized the ICC(3,1) and from now on we refer to this pipeline as ICC.

Power Analysis

Power analysis involves determining the number of subjects needed to show a given effect size, based on the variability of the data (Whitley et al. 2002). An example of such an estimation, can be expressed as

$$\hat{n}_{j,k} = \left(\frac{1.96 \times \sigma_{j,k}}{E_{j,k}}\right)^2$$

where $\hat{n}$ is the number of subjects needed to show an effect E, 1.96 corresponds to a 95% confidence interval, and σ is the group-level standard deviation i.e. the BSV. We estimated the average number of subjects needed to show an effect of either 5% or 10% change in BP_{ND} based on the previously estimated between-subject variabilities for 10 and 20 subjects, including a 95% confidence interval. The effects of either 5% or 10% were estimated as the percent change from the average BP_{ND} for a given region j and pipeline k. Outliers (BP_{ND}'s that were ≤ 0 or ≥ 10) were excluded in the estimation of both σ and E. We note, that there are different ways to estimate the needed sample size depending on the experimental setup, however, as we are mainly interested in the relative differences in sample size between pipelines, this procedure should be sufficient.

RESULTS

An optimal pipeline across subjects and regions (Figure 3)

The evaluation of a median-rank profile for relative pipeline performance for each pipeline and across all subjects (N = 30) and regions (N = 14) is shown in Figure 3, based on the *gSNR*. Higher median rank indicates a higher *gSNR*, and better test-retest performance across subjects and regions for a given preprocessing pipeline. We found a significant pipeline performance effect across subjects ($P < 0.0001$, Friedman test), suggesting the existence of an optimal fixed pipeline. The highest median rank across subjects ($R_{max} = 0.995$), was achieved with the following preprocessing pipeline (FIX): without manual edits in FreeSurfer (FS-RAW), with motion correction (MC), boundary-based co-registration with the time-weighted average image, without partial volume correction (noPVC), and with MRTM2 as preferred kinetic modeling approach. We also identified a subset of several other pipeline choices, that statistically performed equally well as FIX, based on a Dunn-Sidak test, correcting for multiple comparisons for all possible pairwise combinations (P = 0.05). The horizontal dotted line in Figure 3 indicates that pipelines below this line are significantly different from FIX ($R_{cut\text{-}off}$ = 0.989). The pipelines above the cut-off are not significantly different from each other (R_{min} = 0.857).

MC was the factor that influenced pipeline performance most; it consistently increased the median rank when applying MC. The effect of MC also depended on which kinetic model was subsequently applied: whereas the rank for MRTM2 with either MC or nMC were not significantly different from each other (overlapping CI's), the Non-invasive Logan, SRTM and MRTM performed significantly better after MC.

PVC with GTM generally decreased the median rank with increasing PSF (i.e. 0 mm, 2 mm, 4 mm), but the effects were most evident when MC was applied (larger step sizes).

The application of GTM with all PSF options and in combination with MRTM2 showed comparable performance to the noPVC, whereas PVC combined with other kinetic models significantly lowered the test-retest performance, as measured by the *gSNR*.

The co-registration with the time-weighted PET image for both BBR and NMI marginally outperformed the average image, when no MC was applied. This was particularly evident for the higher-rank cases where either Non-invasive Logan or MRTM2 were applied. Only minor effects of co-registration were evident when MC was applied.

The three different FreeSurfer approaches to delineate brain regions did not cause any consistent differences in median rank performance.

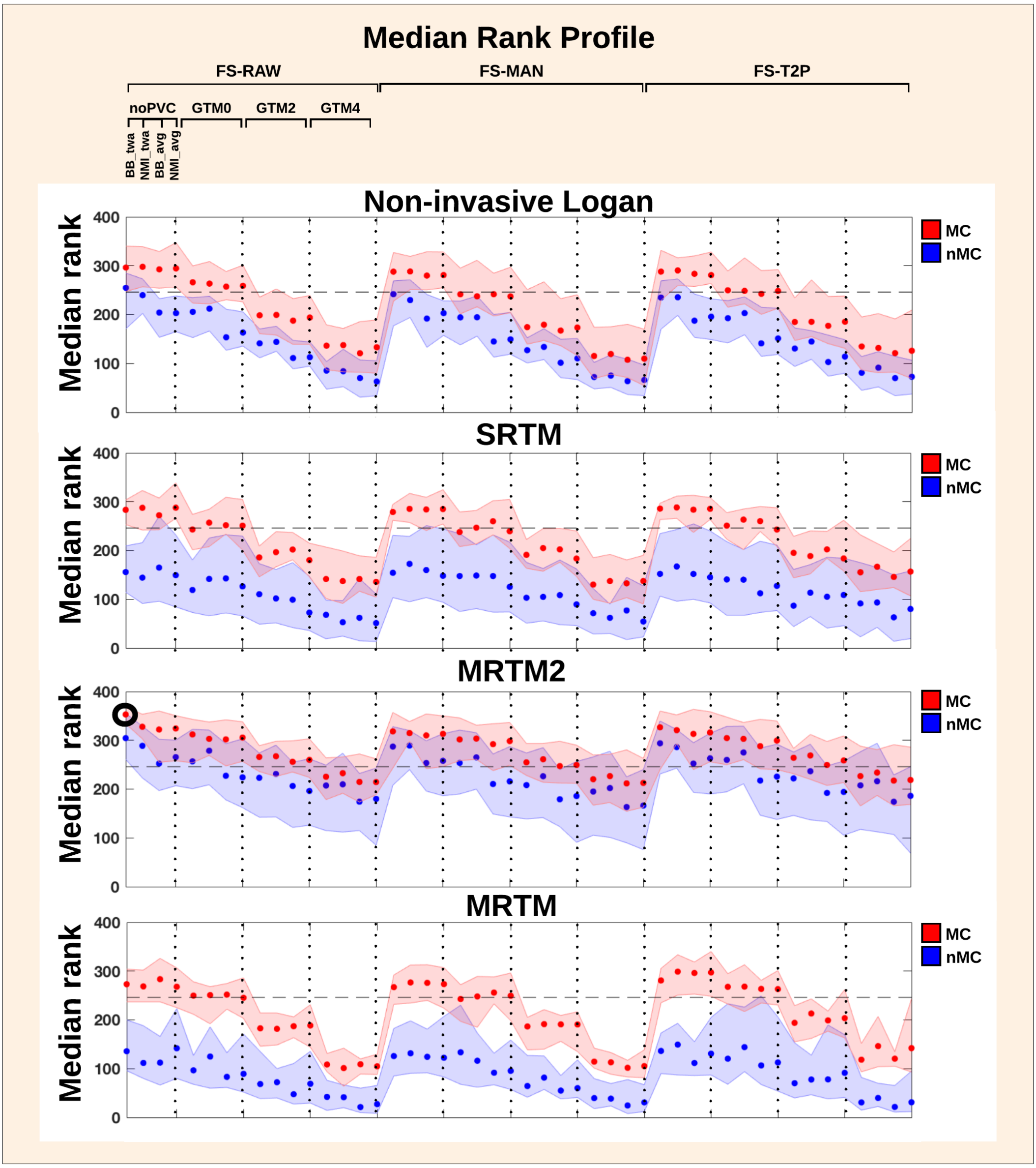

Figure 3: Median rank profile for all pipelines across all subjects. The shaded errorbars indicate 95% confidence intervals. The optimal pipeline across subjects and regions (FIX) is visualized by the black bold circle. The horizontal dotted line indicates that pipelines below this line are significantly different from FIX. The pipelines above the cut-off are not significantly different from each other.

Test-retest bias (Figure 4)

98% of the pipelines revealed a negative test-retest bias (range: -6% to 0%), meaning that the regional BP_{ND}'s were lower on the second scan compared to the first scan. Motion correction had only minor effects on the mean bias for the high-binding regions thalamus, putamen and caudate ranging from 1-2% (Figure S1).

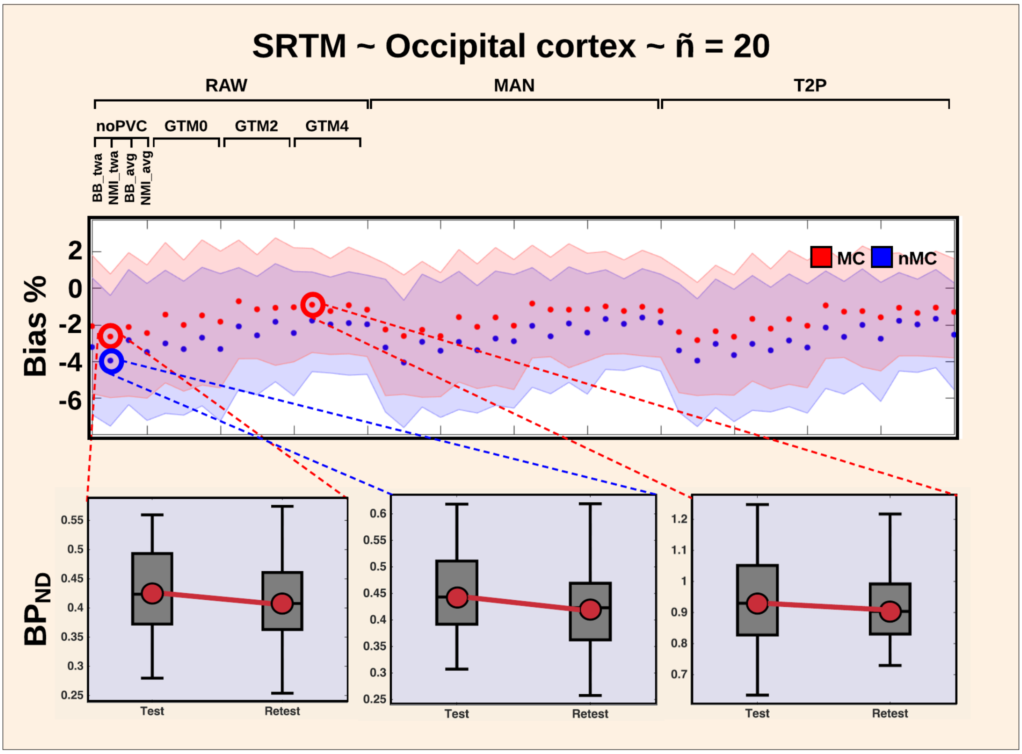

Figure 4: Test-retest bias (%) as a function of pipeline for the occipital cortex, when SRTM is applied. The use of motion correction generally decreases the bias, and ranges from -1% to -4%. This is highlighted by three plots in the bottom, showcasing the test-retest effect on BP_{ND}.

In contrast, when applying SRTM to the occipital cortex, the bias was reduced to -2% when using MC, whereas it was -4% without MC (Figure 4). The bias for the occipital cortex was reduced to -1% when combining MC and GTM with either 2 or 4 mm. For the superior frontal cortex and entorhinal cortex, MRTM and SRTM created some spurious outliers producing up to 15% bias. The amygdala created the highest consistent test-retest bias over all pipelines for both MRTM and SRTM, ranging from -4% when MC was applied, to -6% when no MC was applied (Figure S2).

The amygdala bias was reduced to negative 1-2% when either Non-invasive Logan or MRTM2 was used.

Figures of all estimated biases as a function of sample size ($\tilde{n} = 10$ or $\tilde{n} = 20$), region and preprocessing pipeline are available through the CIMBI database (Knudsen et al. 2016).

Tradeoff in within- and between subject variability at the group level (Figure 5)

The within- and between subject variability were assessed for four different optimization schemes according to the pipelines that for 20 subjects (subsampled 1000 times without replacement) and region *j*, (1) minimized the between-subject variability (BSV), (2) minimized the within-subject variability (WSV) across test and retest, (3) maximized the ICC(3,1) across test and retest, and (4) the fixed optimal pipeline (FIX).

Figure 5 shows the between-subject variability as a function of within-subject variability for these four pipelines, depicted for subcortical and cortical regions. For all regions, we observed a trade-off in between- and within-subject variability, meaning that e.g. minimization of within-subject variability increased between-subject variability, and vice versa (Figure 5). The worst case was the hippocampus showing stable WSV across the pipelines WSV, BSV, FIX and ICC at 10-11%, whereas between-subject variability decreased from 22% to 17% when using the BSV pipeline instead of the ICC pipeline. This translates into 30 fewer subjects needed in a group analysis to detect a 5% difference in BP_{ND} and to achieve similarly powered statistical results (approximately 6 more subjects needed per % increase in BSV).

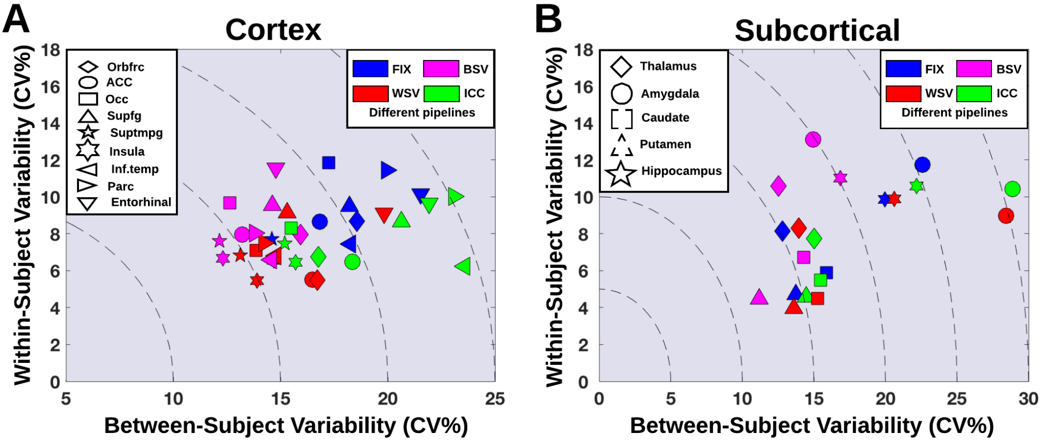

Figure 5: Between-subject variability as a function of within-subject variability for different pipeline optimization schemes, and for both cortical (A) and subcortical regions (B).

Table 1 lists the optimal preprocessing pipelines for the 14 regions. Across all regions, use of either MRTM or MRTM2 consistently minimized WSV (Table 1). Notably, the application of GTM 4 mm minimized the between-subject variability in all regions except for the amygdala, thalamus and hippocampus, and the within-subject variability was similarly minimized in all cortical regions, except from the regions insula and entorhinal cortex (Table 1).

Table 1: Overview of optimal pipelines for the brain regions amygdala, thalamus, putamen, caudate, anterior cingulate, hippocampus, orbital frontal cortex (FC), occipital, superior FC, superior temporal gyrus (TG), insula, medial-inferior TG, parietal cortex and entorhinal, when optimized by median-rank (FIX), within-subject variability (WSV), between-subject variability (BSV) and intra-class correlation (ICC). 1st letter (Delineation of regions; A=FS-raw, B=FS-man, C=FS-T2p), 2nd letter (Motion Correction (MC); A=MC, B=noMC), 3rd letter (Co-registration; A=BB_{TWA}, B=NMI_{TWA}, C=BB_{AVG}, D=NMI_{AVG}), 4th letter (Partial Volume Correction (PVC); A=noPVC, B=Geometrix Transfer Matrix (GTM) 0 mm, C=GTM 2 mm, D=GTM 4 mm), 5th letter (Kinetic modeling; A=MRTM, B=MRTM2, C=SRTM, D=Non-invasive Logan).

	FIX	**WSV**	**BSV**	**ICC**
Amygdala	AAAAB	CBBCB	BAAAD	ABACB
Thalamus	AAAAB	BAAAA	ABBAD	BABDA
Putamen	AAAAB	CAAAA	CADDA	AABDA
Caudate	AAAAB	CAADB	CADDA	AAADB
Anterior Cingulate	AAAAB	BBADB	ABDDD	CBADB
Hippocampus	AAAAB	BBBAB	ABBAD	CBBCB
Orbital FC	AAAAB	BBBDB	CBDDD	BBADB
Occipital	AAAAB	BABDB	ABDDC	CABDA
Superior FC	AAAAB	ABCDB	ABBDA	CBADC
Superior TG	AAAAB	BBBDB	AABDD	BBABB
Insula	AAAAB	CABBA	BABDD	CBBDB
Medial-Inferior TG	AAAAB	BBBDB	BABDB	CBBDB
Parietal C	AAAAB	ABADA	ABCDB	BBABC
Entorhinal	AAAAB	CABAB	CBBDD	BABDB

Figure 6A and 6B display the within- and between-subject variability as a function of region, and with or without application of MC. In Figure 6C and 6D, the within- and between subject variability is displayed for noPVC vs. GTM 4 mm. Figure 6 shows the impact on within- (E) and between-subject (F) variability of choosing SRTM versus MRTM2.

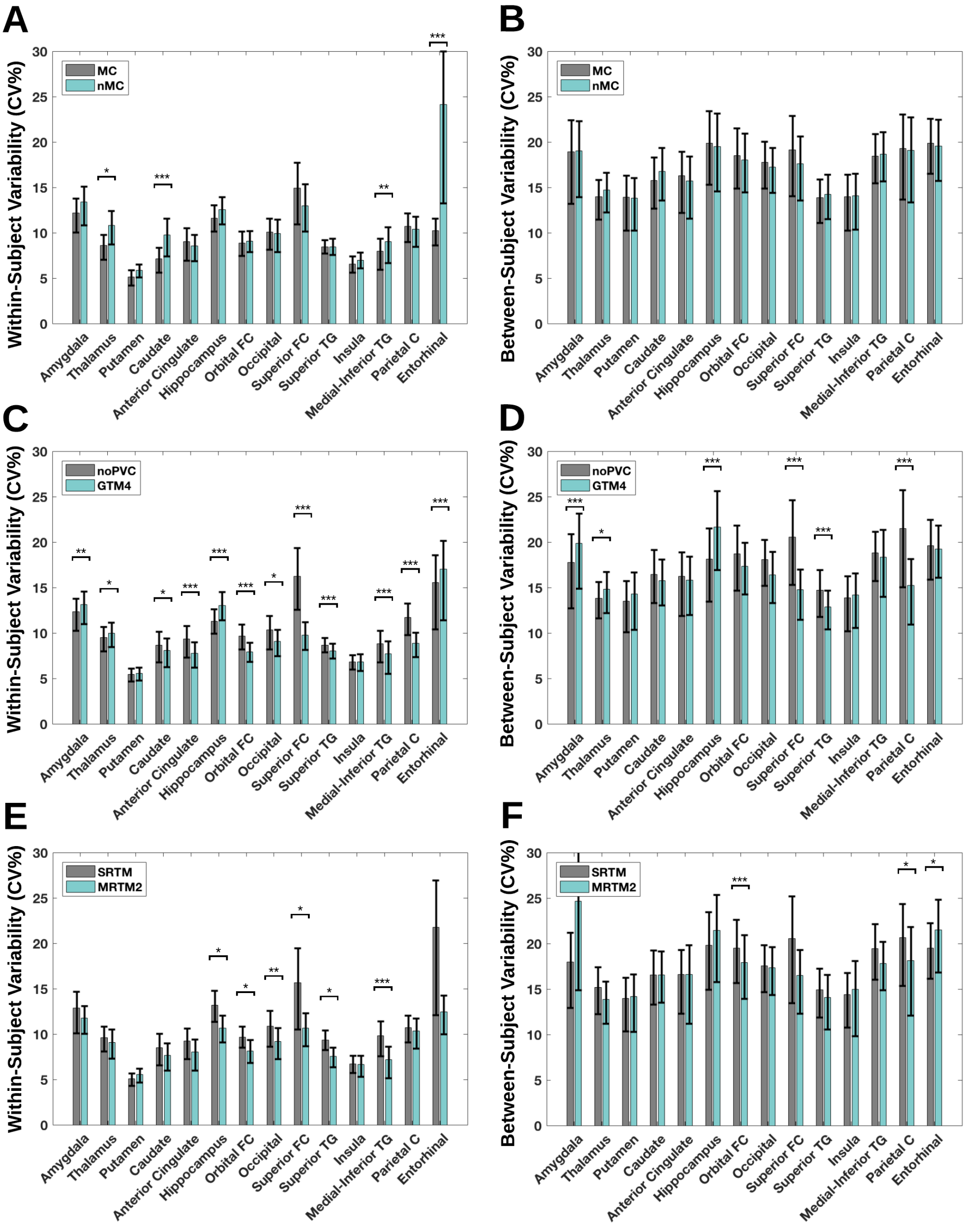

Figure 6: (A) within-subject variability for 14 regions with or without motion correction, including a 95% confidence interval (B) between-subject variability for 14 regions with or without motion correction, including a 95% confidence interval (C-D) similar to A and B, but with either no partial volume correction (noPVC) or with the Geometric Transfer Matrix (GTM) and a point spread function assumption of 4 mm (E-F) similar to A and B, but with the application of either the Simplified Reference Tissue Model (SRTM) or

the Multilinear Reference Tissue Model 2 (MRTM2) as kinetic modeling choice. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, FDR corrected for multiple comparisons (FDR=0.05).

Figures of all estimated variabilities as a function of sample size ($\tilde{n} = 10$ or $\tilde{n} = 20$), region and preprocessing pipeline are avilable through the CIMBI database (Knudsen et al. 2016).

Power analysis across preprocessing pipeline choices (Figure 7)

The effects of pipeline choice on the sample size required to show a given effect size were also examined. Figure 7 shows the sample size required to detect a 5% difference from the anterior cingulate mean BP_{ND}. A 95% CI is plotted for all pipeline combinations for Non-invasive Logan, SRTM and MRTM2 for $\tilde{n} = 20$.

The greatest reduction in sample size was seen with the choice of kinetic modeling: the Non-invasive Logan in combination with GTM 4 mm and no MC was associated with a sample size of 27 subjects [CI: 16 subjects – 36 subjects] (Figure 7). Notably, when combined with GTM 4 mm, the FS-T2P stream resulted in substantially higher sample size (Figure 7). This result was driven by a single subject producing a substantially higher BP_{ND} after FS-T2P correction, consequently increasing the between-subject variability.

Figures for sample size as a function of pipelines, regions, subjects and effect sizes are available through the CIMBI database (Knudsen et al. 2016).

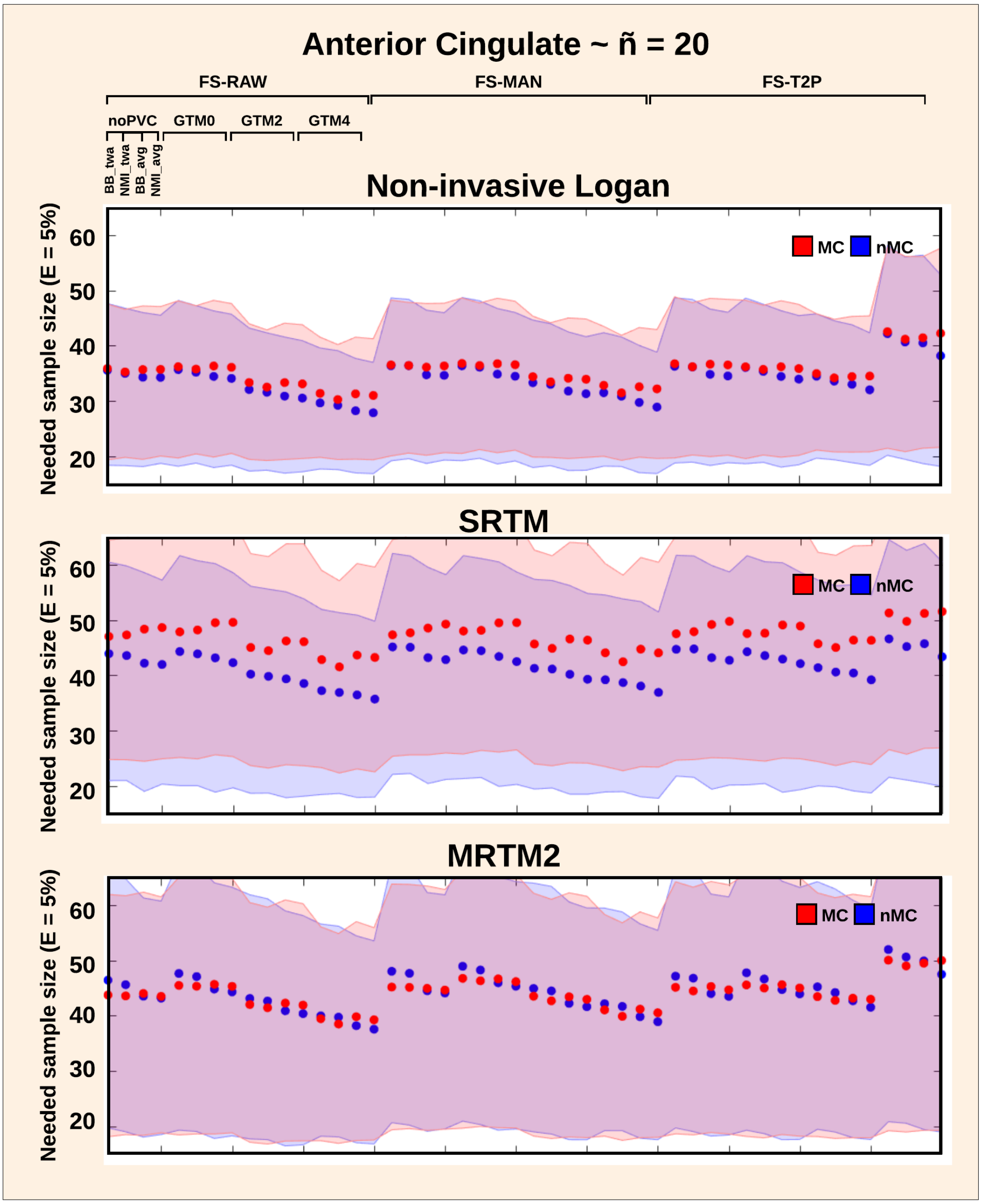

Figure 7: Sample size required to detect a group differece of 5% in BP_{ND} for the anterior cingulate cortex, depending on the kinetic modeling approach (Non-invasive Logan, SRTM and MRTM), and for all other pipeline choices. The blue is without motion correction (nMC) and the red line is with motion correction (MC). The shaded error bars indicate the 95% confidence interval, estimated by randomly choosing 20 subjects over 100 resampling's.

DISCUSSION

In a comprehensive preprocessing framework, we report the evaluation of the impact of 384 different preprocessing pipelines on a set of common performance metrics, based on test-retest [11C]DASB neuroimaging data. Our findings suggest that the observed complex interaction between various preprocessing steps and brain regions necessitates careful consideration of the performance of a chosen preprocessing pipeline and the final outcome measure BP_{ND}.

Test-retest bias on [11C]DASB binding

Whereas test-retest studies are generally considered to provide valuable information about the repeatability of PET measures, variability may not only arise from measurement errors but also from biological variations between scans. Independent of the chosen preprocessing pipeline, we consistently found lower [11C]DASB BP_{ND} at the second compared to the first scan.
This observation was also made by Kim et al. 2006 who reported a negative bias of 2.5% - 7.5% between first and second scan. Two other test-retest studies by Frankle et al. 2005 and Ogden et al. 2006 did not apply any test-retest bias metric in their evaluation.
Regardless, if the negative test-retest bias is a true biological effect or if it is introduced in the data-acquisition and/or preprocessing stage, care must be taken in analysis of longitudinal data to avoid attributing an effect to a treatment/condition that was actually due to the retest alone. Further, test-retest studies with a biologically determined bias means that attempts to define a pipeline that minimizes the bias may be counterproductive.

Extensive research in humans support a number of factors affecting cerebral [11C]DASB binding. Diurnal variation has been reported to affect 5-HTT binding (Meyerson et al. 1989) causing an increase in measurement variability if test and retest scans are executed in the morning and afternoon. The effects of having repeated tracer injections may introduce carry-over effects, or induce internalization or conformational changes of the 5-HTT to a different state (Zhang et al. 2016). The data used in the present study were acquired with an interval of 5 weeks, at the same phase in the menstrual cycle and at the same time of the day, which makes it unlikely that carry-over effects, hormonal or diurnal changes explain our observation. Kim et al. 2006 also discussed the possibility of increased stress levels at the first scan, elevating the circulation of cortisol, consequently increasing 5-HTT synthesis and thereby potentially lowering 5-HTT binding observed at retest. While increased stress levels at the first scan may be causing the negative bias, it may also be attributed to a change in levels of motion between test and retest, with less motion contributing to an increase in SNR (if subjects are more calm at retest). In contrast, high levels of motion may have substantial impact on the scanner reconstruction, potentially either over- or underestimating the true uptake by affecting attenuation- and scatter correction (Van den Heuvel et al. 2013).

Irrespective whether the bias appears from biological variations and/or is caused by data acquisition and/or preprocessing, it should be taken into account if it is likely to have an impact on the scientific question. More specifically, scientific question could depend on e.g. 1) region 2) one or more scans 3) structural abnormalities such as atrophy, and 4) disorder related head motion.

In contrast, bias may trade off with true between subject variability where greater variability may reflect more accurate between-subject biological variation. This is contrary to focusing only on group mean differences which we have mainly focused on in this study.

Nevertheless, depending on the ability to remove potential biases and depending on the size of their individual contributions, both the within- and between-subject variability should subsequently be assessed to decide whether an estimated effect can be considered a true biological effect or not.

Impact of preprocessing pipeline strategy

Using our evaluation framework, we identified a set of optimal pipelines across subjects and regions showing significant effects for MC, co-registration, PVC and kinetic modeling (Figure 3). Although it is well-known that MC can have a significant impact on PET results (Montgomery et al. 2006), about 40% of published [^{11}C]DASB PET neuroimaging studies leave out MC (Nørgaard et al. 2018).

One of the most consistent outcomes of our analysis was that MC had an impact on the pipeline performance. Given that we only included scans with < 3 mm median movement, MC is likely to have an even larger impact in people with larger head motion. Conversely, in the absence of motion, MC will lead to some degree of smoothing due to interpolation which might have improved the performance. The improved performance with MC could also result from higher noise-levels at the end of the scan where the distribution of radioactivity is lower producing less true counts, or by a re-distribution of the tracer. Freire and Mangin 2001, and Orchard and Atkins 2003, demonstrated that least-squares cost functions may be susceptible to fMRI activation biases, which for PET means that the MC algorithm may attempt to incorrectly account for motion if the VOI has low SNR, or if the tracer distribution in the target volume changes significantly over time compared to the reference volume. While we identified an overall impact of MC on pipeline ranking using the performance metric gSNR (Figure 3), we also found that particularly the thalamus, caudate, medial-

inferior TG and entorhinal cortex (Figure 6A) contributed to the within-subject variability. This is important, because thalamus and caudate are often used as high-binding regions in MRTM2, affecting the estimation of k_2' and consequently the BP_{ND}. While choice of reference region will impact all RTM's, MRTM2 and Non-invasive Logan are particularly sensitive to the choice of an adequate high-binding region for estimation of k_2' (Ichise et al. 2003, Mandeville et al. 2016). In previous studies, different high-binding regions have been used, without any particular justification, e.g., raphe, thalamus and striatum (Kim et al. 2006); midbrain, thalamus and striatum (Matsumoto et al. 2010); raphe (Hesse et al. 2011); occipital cortex (Brown et al. 2007); or thalamus (James et al. 2017). The same group may even choose different high-binding regions across studies, e.g. Gryglewski et al. 2017 (striatum) and James et al. 2017 (thalamus), and some studies do not mention which high-binding region was chosen (e.g. Zientek et al. 2016). A few studies also cite other studies as justification for using a high-binding region, but then use another high-binding region than the cited study (e.g. Kupers et al. 2010 & Frokjaer et al. 2009).
Based on previous literature (e.g. Beliveau et al. 2016), we rather arbitrarily decided to use thalamus and striatum as high-binding regions for estimation high-binding regions for estimation of k_2'. However, as displayed in Figure 6, this choice may not be optimal, as the putamen not only minimizes the within- and between-subject variability relative to thalamus and caudate, it is also the region least affected by preprocessing strategy; MC, PVC and kinetic modeling. The putamen delineation in FreeSurfer is a more homogeneous gray-matter region compared to thalamus (see supplementary text 3 for evaluation), and does not suffer from the same severe partial volume effects as caudate does because of its proximity to CSF. Therefore, one could consider the putamen to be the optimal choice of high-binding region to minimize potential biases originating from subject-dependent differences. In a post-hoc test we evaluated the use of putamen as high-binding region, and this indeed lowered the between-subject variability with 1-10% depending on the VOI, at the expense of bias in group mean. As this post-hoc test is a circular analysis, our observation needs to be tested in an independent cohort.

The performance of the optimal pipeline was also largely dependent on the use of noPVC or GTM with either 0 mm, 2 mm or 4 mm, with the latter contributing negatively to the overall pipeline rank, as highlighted by the performance metric gSNR (Figure 3). At first sight, this effect would seem to be caused by violations of the GTM assumptions, presumably the PSF and the constant uptake within each VOI. For subcortical regions, the thalamus delineation and consequently the [^{11}C]DASB uptake homogeneity has been shown to vary substantially between atlases (Nørgaard et al. 2015), which may make the estimate more noisy. For cortical regions where the 5-HTT density is relatively low and the average cortical thickness only 3 mm (Fischl et al. 2000), the voxel-wise noise level may be higher than in the subcortical regions.

However, in a post-hoc analysis on variability (Figure 6C and 6D), we identified a distinct difference in PVC performance across subcortical and cortical regions. While the application of GTM 4 mm caused a significant decrease in within-subject variability in all cortical regions except for the insula and entorhinal cortex, it significantly increased it in the amygdala, thalamus and hippocampus. More specifically, the amygdala and hippocampus were critically affected by this preprocessing step, increasing both within-subject and between-subject variability. This may be attributed to partial volume effects being similar in the amygdala, hippocampus, and cerebellum, resulting in more unstable estimates due to the dependencies between regions.

Regardless of the contribution to noise of the GTM, PVC is still highly recommended in studies where brain atrophy interacts with an effect of interest (e.g., age or diagnosis). Failure to properly account for partial volume effects in these cases can falsely inflate or degrade the effect of interest (Greve et al. 2016).

Amygdala and hippocampus have medium to high 5-HTT density and with long uptake times, the TACs tend to reflect irreversible binding, which may compromise the identification of stable model parameters, resulting in noisy estimates. For the pipeline-rank performance metric (Figure 3), if within-subject variability increases when GTM 4 mm is applied, this will have a negative impact on the pipeline-rank metric, as it largely depends on the Pearson's correlation coefficient.

This variability was reduced after MC, but the remaining difference in pipeline performance was significantly affected by the choice of kinetic modeling. Depending on the difference in noise-levels at test and retest due to e.g. motion, this may be caused by a bias in the BP_{ND} estimates from kinetic models using non-invasive Logan, SRTM and MRTM, consequently reducing the test-retest performance. This is because BP_{ND} estimates from kinetic models are subject to noise-dependent bias, meaning that as the noise-level increases, the estimated BP_{ND} deviates from the true value (Ichise et al. 2003). The MRTM2 has no noisy term as independent variable when fitting the kinetic model parameters with multi-linear regression, thus effectively reducing the noise-induced bias and improving overall performance (Ichise et al. 2003).

Trade-off in within- and between-subject variability at the group level

The within-subject and between-subject variability analysis revealed important trade-offs in pipeline performance as a function of region (Figure 5). Minimization of between-subject variability increased within-subject variability relative to the fixed pipeline, particularly for the amygdala, thalamus and hippocampus quantified with the non-invasive Logan model. Quantification with the non-invasive Logan method is often preferred due to it having the lowest between-subject coefficient of variation (Tyrer et al. 2016, Logan et al. 1996), however, our analyses indicate that this comes at the expense of a 3-5% increase in within-subject variability (range: 10% – 14%), as shown in Figure 5. Consequently, depending on the experimental setup (i.e. group or longitudinal study) the choice of preprocessing should be selected with caution and consideration of the study goals and design.

The analysis on the effects of spatial normalization on BSV and WSV (supplemental material), showed only a small difference in terms of average BP_{ND} compared to without normalization, but the within- and between-subject variability were substantially increased by a factor of 2-4. This effect is likely to be caused by contamination of CSF and white-matter in the VOI in standard space, requiring a substantial increase in number of subjects needed to obtain similar statistical power.

The within-subjects design captures the difference among conditions (i.e., test and retest) and has the clear advantage that fewer subjects are required. However, the within-subjects design is subject to learning effects across conditions if the design is not placebo vs active, which is not the case for between-subject designs. Care must therefore be taken in the analysis of longitudinal data to avoid attributing an effect to a treatment/condition that was actually due to a potential retest bias.

In the absence of a “ground truth”, it remains a challenge to select the optimal preprocessing pipeline, and it may take alternative performance metrics to quantitatively evaluate and compare various pipelines (Strother et al. 2002, Churchill et al. 2015). We want to emphasize that the aim of this study was not to identify a definitive preprocessing pipeline for $[^{11}C]$DASB data, but to quantify the impact of the preprocessing choices selected in this study and their uncertainty on BP_{ND}.

Enhancement of study power with optimal preprocessing pipelines

The comparison of subjects needed to show a given effect size, provided insight into the effect of preprocessing pipeline choice on sample size and as a function of region, based on the between-subject variability performance metric. The test-retest studies published so far for

[^{11}C]DASB included between 8 and 11 volunteers (Ogden et al. 2007, Frankle et al. 2006, Kim et al. 2005) (the present study includes 30 subjects) and sample sizes in published [^{11}C]DASB PET studies range from 5 (Ogawa et al. 2014) to 83 subjects (Miller et al. 2013), but with approximately 20 subjects being the most common (Nørgaard et al. 2018). However, while the sample size required to show an effect should ultimately be determined by the variability of the measured random variable (i.e. BP_{ND}), power analyses may become biased if incorrect variability measures are used. Therefore, here we provide an estimate of what sample size is needed to show an effect of either 5% og 10% difference in BP_{ND} as a function of pipeline choice and for a specific hypothesis related to a given region (available through the CIMBI database (Knudsen et al. 2016)).

As highlighted previously, there exist a trade-off between the optimization of within- and between subject variation as a function of VOI and preprocessing pipeline. This ultimately affects our recommendation of preprocessing strategy to maximize power. For example, given no apriori hypothesis related to a specific region, we recommend the pipeline from the rank analysis using gSNR as performance metric. However, as the gSNR metric is mostly sensitive to within-subject variance and because the power estimation is largely driven by between-subject variance, there will exist other pipelines that maximizes power by minimizing between-subject variance at the expense of increased within-subject variance.

We strongly suggest that researchers take the reported biases and variations into account when they conduct power analyses prior to a study. In addition, we recommend choosing a fixed preprocessing pipeline prior to data acquisition depending on the researcher's biological question, as this should help to avoid underpowered studies.

While it is quite common in the PET community to perform regional analyses, several attempts have also been extended to both voxel- and surface based analyses. The effects of bias and variance trade-offs as a function of various tracers and preprocessing pipelines are thus largely unknown for these types of analyses, and only a few papers have attempted to address some of these challenges for PET (e.g. Greve et al. 2014) and fMRI (e.g. Churchill et al. 2015).

All the reported results and analyses are available through the CIMBI database (Knudsen et al. 2016).

LIMITATIONS AND FUTURE CONSIDERATIONS

Our study is not without limitations. The results were derived from the radiotracer [^{11}C]DASB measured in the HRRT scanner; however, we expect the results to generalize to other radiotracers and scanners, with a possible exception of PVC. Inclusion of PVC only had minor effects on most performance metrics. While this may be a specific finding in the context of using the HRRT with the PSF-OSEM reconstruction, it may be questionable whether PVC would generably be favourable in cases of PET images obtained with a conventional PET scanner offering a resolution of 4-5 mm. However, since the the radioligand kinetic behaviour and the spatiotemporal noise will differ between radioligands, separate assessements of each radioligand's pipeline performace may be warranted as part of new validation papers.

We observed only minor differences in performance between FS-RAW, FS-MAN and FS-T2P, but substantial differences in regional binding estimates could presumeably be obtained if a different brain atlas (e.g. PVElab or AAL) is used. We deliberately abstained from testing other segmentation atlases since the outcome was likely to be influenced by differences in volumes, etc. FreeSurfer returns 41 regions per hemisphere and to make the results more comparable to other atlases, we chose to extract only a subset of 14 regions covering all major parts of the brain.

The results and interpretations of this study can therefore not be generalized to the remaining 27 regions, with the tradeoff being that the results are more comparable to regions from other atlases. Furthermore, the merging of regions between hemispheres is also a limitation if lateralized effects are present. On the other hand, averaging across hemispheres is commonly done in PET studies because it reduces the number of statistical tests.

However, even though the extent to which a change in brain atlas affects regional binding is quantifiable, it is not trivial to determine whether a decrease/increase in the performance metrics suggests a better choice of atlas. Further investigation is therefore needed in order to understand this question.

With respect to kinetic modeling we decided not to include SRTM2 because SRTM2 and MRTM2 perform similarly well. Optimally, the current framework should be expanded to include data from different scanners and other acquisition parameters to evaluate inter-site differences, however data sharing initiatives are needed to accomplish this task (Knudsen et al., 2016). This is beyond the scope of this paper.

Last but not least, the chosen steps of preprocessing in this study is also a limitation. Our future goal is to make the data publicly available, so researchers can download the data and benchmark their own preprocessing pipeline using the same performance metrics and the same data. The results of the benchmark can subsequently be made publicly available on a website or in a database. The effort aligns well with current interests in the PET community, as was highlighted at the NRM2018 PET Grand Challenge (www.petgrandchallenge.com).

CONCLUSION

In summary, we provide evidence that preprocessing pipeline choices have significant impact on [^{11}C]DASB BP_{ND} in a distributed set of brain regions, as evaluated by 7 performance metrics. Given that *no* apriori hypothesis exist, we recommend researchers use the FIX pipeline (with MC, co-registration BBR and the time-weighted PET image, no PVC, and kinetic modeling using MRTM2). Given a specific clinical hypothesis (e.g. change in binding in putamen), we recommend researchers to use Table 1 as a guideline, with longitudinal studies using the WSV column, as this measure ensures minimum test-retest variability between scan sessions. For cross-sectional studies, we recommend researchers choose a pipeline that minimizes both within- and between subject variability (i.e. either the BSV or ICC column in Table 1), as this should ensure a compromise between low within-subject variability and low between-subject variability.

The heterogeneity of pipeline effects among the evaluated young and healthy subjects, emphasizes the relative importance of pipeline performance and that the preprocessing pipeline should be selected with great caution.

The presented evaluation framework can easily be expanded to include more pipelines and different data, but this would come at the expense of increased computational time (combinatorial explosion) and proper evaluation strategies.

To conclude, these findings provide novel information of what can be expected of variability in either previous or future [^{11}C]DASB studies, given a specific hypothesis related to i.e. region, sample size, and preprocessing pipeline choice.

ACKNOWLEDGMENTS

We wish to thank all the participants for kindly joining the research project. We thank the John and Birthe Meyer Foundation for the donation of the cyclotron and PET scanner. Peter Steen Jensen, Vincent Beliveau and Patrick M. Fisher are gratefully acknowledged. Finally, we express our gratitude to the anonymous reviewers for the many suggestions that helped improve this paper.

FUNDING

MN was supported by the National Institutes of Health (Grant 5R21EB018964-02), the Lundbeck Foundation (Grant R90-A7722), and the Independent Research Fund Denmark (DFF-1331-00109 & DFF-4183-00627).

DISCLOSURE/CONFLICT OF INTEREST

The authors declare no conflict of interest or financial disclosures. SCS is the consulting Chief Scientific Officer at ADMdx, Inc.

REFERENCES

Best SE, Sarrel PM, Malison RT, Laruelle M, Zoghbi SS, Baldwin RM, Seibyl JP, Innis RB, van Dyck CH. Striatal dopamine transporter availability with [123I]beta-CIT SPECT is unrelated to gender or menstrual cycle. Psychopharmacology (Berl). 2005 Dec;183(2):181-9.

Boileau I, Warsh JJ, Guttman M, et al. Elevated serotonin transporter binding in depressed patients with Parkinson's disease: a preliminary PET study with [11C]DASB. Mov Disord. 2008 Sep 15;23(12):1776-80.

Brown AK, George DT, Fujita M, Liow JS, Ichise M, Hibbeln J, Ghose S, Sangare J, Hommer D, Innis RB. PET [11C]DASB imaging of serotonin transporters in patients with alcoholism. Alcohol Clin Exp Res. 2007 Jan;31(1):28-32.

Cannon DM, Klaver JM, Klug SA, Carlson PJ, Luckenbaugh DA, Ichise M, Drevets WC. Gender-specific abnormalities in the serotonin transporter system in panic disorder. Int J Neuropsychopharmacol. 2013 May;16(4):733-43.

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., … Strother, S. C. (2012). Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Human Brain Mapping*, *33*(3), 609–627.

Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., & Strother, S. C. (2015). An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. *PLoS ONE*, *10*(7), 1–25.

Fischl B. FreeSurfer. Neuroimage. 2012 Aug 15; 62(2): 774-781.

Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci U S A. 2000 Sep 26;97(20):11050-5.

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM. Cereb Cortex. 2004 Jan;14(1):11-22.

Frankle, W. G., Slifstein, M., Gunn, R. N., Huang, Y., Hwang, D. R., Darr, E. A., Narendran, R., Abi-Dargham, A., and Laruelle, M. (2006). Estimation of serotonin transporter parameters with 11C-DASB in healthy humans: reproducibility and comparison of methods. J Nucl Med, 47:815–826.

Frick, A., Åhs, F., Engman, J., Jonasson, M., Alaie, I., Björkstrand, J., Frans, Ö., Faria, V., Linnman, C., Appel, L., Wahlstedt, K., Lubberink, M., Fredrikson, M., and Furmark, T. (2015). Serotonin Synthesis and Reuptake in Social Anxiety Disorder. JAMA Psychiatry, (JUNE):E1–E9.

Frokjaer, V. G., Erritzoe, D., Holst, K. K., Jensen, P. S., Rasmussen, P. M., Fisher, P. M., Baaré, W., Madsen, K. S., Madsen, J., Svarer, C., and Knudsen, G. M. (2013). Prefrontal serotonin transporter availability is positively associated with the cortisol awakening response. European Neuropsychopharmacology, 23(4):285–294.

Frokjaer, V. G., Pinborg, A., Holst, K. K., Overgaard, A., Henningsson, S., Heede, M., Larsen, E. C., Jensen, P. S., Agn, M., Nielsen, A. P., Stenbæk, D. S., Da Cunha-Bang, S., Lehel, S., Siebner, H. R., Mikkelsen, J. D., Svarer, C., and Knudsen, G. M. (2015). Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: A positron emission tomography study. Biological Psychiatry, 78(8):534–543.

Frokjaer, V. G., Vinberg, M., Erritzoe, D., Svarer, C., Baaré, W., Budtz-Joergensen, E., Madsen, K., Madsen, J., Kessing, L. V., and Knudsen, G. M. (2009). High familial risk for mood disorder is associated with low dorsolateral prefrontal cortex serotonin transporter binding. NeuroImage, 46(2):360–366.

Ganz, M., Feng, L., Hansen, H. D., Beliveau, V., Svarer, C., Knudsen, G. M., and Greve, D. N. (2017). Cerebellar heterogeneity and its impact on PET data quantification of 5-HT receptor radioligands. Journal of Cerebral Blood Flow & Metabolism, page 0271678X1668609.

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.

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

Greve, D. N., Salat, D. H., Bowen, S. L., Izquierdo-Garcia, D., Schultz, A. P., Catana, C., Becker, J. A., Svarer, C., Knudsen, G. M., Sperling, R. A., and Johnson, K. A. (2016). Different partial volume correction methods lead to different conclusions: An 18F-FDG-PET study of aging. NeuroImage, 132:334–343.

Greve, D. N., Svarer, C., Fisher, P. M., Feng, L., Hansen, A. E., Baare, W., Rosen, B., Fischl, B., and Knudsen, G. M. (2014). Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data. NeuroImage, 92:225–236.

Gryglewski G, Rischka L, Philippe C, Hahn A, James GM, Klebermass E, Hienert M, Silberbauer L, Vanicek T, Kautzky A, Berroterán-Infante N, Nics L, Traub-Weidinger T, Mitterhauser M, Wadsak W, Hacker M, Kasper S, Lanzenberger R. Simple and rapid quantification of serotonin transporter binding using [^{11}C]DASB bolus plus constant infusion. Neuroimage. 2017 Jan 22;149:23-32.

Hesse S, Stengler K, Regenthal R, Patt M, Becker GA, Franke A, Knüpfer H, Meyer PM, Luthardt J, Jahn I, Lobsien D, Heinke W, Brust P, Hegerl U, Sabri O. The serotonin transporter availability in untreated early-onset and late-onset patients with obsessive-compulsive disorder. Int J Neuropsychopharmacol. 2011 Jun;14(5):606-17.

Houle S, Ginovart N, Hussey D, Meyer JH, Wilson AA. Imaging the serotonin transporter with positron emission tomography: initial human studies with [11C]DAPP and [11C]DASB. Eur J Nucl Med. 2000 Nov;27(11):1719-22.

Ichise, M., Liow, J.-S., Lu, J.-Q., Takano, A., Model, K., Toyama, H., … Carson, R. E. (2003). Linearized reference tissue parametric imaging methods: application to [11C]DASB positron emission tomography studies of the serotonin transporter in human brain. *Journal of Cerebral Blood Flow and Metabolism : Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *23*(9), 1096–1112.

James GM, Baldinger-Melich P, Philippe C, Kranz GS, Vanicek T, Hahn A, Gryglewski G, Hienert M, Spies M, Traub-Weidinger T, Mitterhauser M, Wadsak W, Hacker M, Kasper S, Lanzenberger R. Effects of Selective Serotonin Reuptake Inhibitors on Interregional Relation of Serotonin Transporter Availability in Major Depression. Front Hum Neurosci. 2017 Feb 6;11:48.

Jovanovic H, Karlsson P, Cerin A, Halldin C, Nordström AL. 5-HT(1A) receptor and 5-HTT binding during the menstrual cycle in healthy women examined with [(11)C] WAY100635 and [(11)C] MADAM PET. Psychiatry Res. 2009 Apr 30;172(1):31-7.

Jovicich J, Czanner S, Greve DN, Haley E, van der Kouwe A, Gollub R, Kennedy D, et al. Reliability in Multi–Site Structural MRI Studies: Effects of Gradient Non–Linearity Correction on Phantom and Human Data. NeuroImage, 2006; 30(2): 436–43.

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.

Kim, J. S., Ichise, M., Sangare, J., and Innis, R. B. (2006). PET Imaging of Serotonin Transporters with [11C]DASB: Test- Retest Reproducibility Using a Multilinear Reference Tissue Parametric

Imaging Method. J. Nucl. Med., 47(2):208–214.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2010). Circular analysis in systems neuroscience – the dangers of double dipping. *Nat Neurosci*, *12*(5), 535–540.

Lanzenberger, R., Kranz, G. S., Haeusler, D., Akimova, E., Savli, M., Hahn, A., Mitterhauser, M., Spindelegger, C., Philippe, C., Fink, M., Wadsak, W., Karanikas, G., and Kasper, S. (2012). Prediction of SSRI treatment response in major depression based on serotonin transporter interplay between median raphe nucleus and projection areas. NeuroImage, 63(2):874–881.

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

Logan J, Fowler JS, Volkow ND, Wang GJ, Ding YS, Alexoff DL: Distribution volume ratios without blood sampling from graphical analysis of PET data. J Cereb Blood Flow Metab 1996, 16(5):834-840.

Madsen K, Haahr M, Marner L, Keller SH, Baaré W, Svarer C, Hasselbalch SG, Knudsen GM. Age and Sex Effects on 5-HT(4) Receptors in the Human Brain – A [11C]SB207145 PET Study. Journal of Cerebral Blood Flow & Metabolism 2011;31(6):1475-81.

Marner, L., Frokjaer, V. G., Kalbitzer, J., Lehel, S., Madsen, K., Baaré, W. F. C., Knudsen, G. M., and Hasselbalch, S. G. (2012). Loss of serotonin 2A receptors exceeds loss of serotonergic projections in early Alzheimer's disease: A combined [11C]DASB and [18F]altanserin-PET study. Neurobiology of Aging, 33(3):479–487.
Matsumoto, R., Ichise, M., Ito, H., et al. (2010). Reduced serotonin transporter binding in the insular cortex in patients with obsessive-compulsive disorder: A [^{11}C]DASB PET study. NeuroImage, 49(1):121–126.

McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J. A., Coman, I. L., & Kates, W. R. (2015).

A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience*, *9*(OCT), 1–18.

Meyer JH, Wilson AA, Ginovart N, Goulding V, Hussey D, Hood K, Houle S. Occupancy of serotonin transporters by paroxetine and citalopram during treatment of depression: a [(11)C]DASB PET imaging study. Am J Psychiatry. 2001 Nov;158(11):1843-9.

Meyerson LR, Strano R, Ocheret D. Diurnal concordance of human platelet serotonin content and plasma alpha-1-acid glycoprotein concentration. Pharma- col Biochem Behav. 1989;32:1043–1047. 32.

Miller, J. M., Hesselgrave, N., Ogden, R. T., Sullivan, G. M., Oquendo, M. A., Mann, J. J., & Parsey, R. V. (2013). Positron emission tomography quantification of serotonin transporter in suicide attempters with major depressive disorder. *Biological Psychiatry*, *74*(4), 287–295.

Montgomery, A. J., Thielemans, K., Mehta, M. A., Turkheimer, F., Mustafovic, S., & Grasby, P.M. (2006). Correction of Head Movement on PET Studies: Comparison of Methods. *J. Nucl. Med.*, *47*(12), 1936–1944.

Nørgaard M, Ganz M, Fisher PM, et al. (2015). Estimation of regional seasonal variations in SERT-levels using the FreeSurfer PET pipeline: A reproducibility study. In: Proceedings of the MICCAI workshop on computational methods for molecular imaging 2015.

Nørgaard M, Ganz M, Svarer C, Feng L, Ichise M, Lanzenberger R, Lubberink M, Parsey RV, Politis M, Rabiner EA, Slifstein M, Sossi V, Suhara T, Talbot PS, Turkheimer F, Strother SC, Knudsen GM. 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.

Ogawa, K., Tateno, A., Arakawa, R., Sakayori, T., Ikeda, Y., Suzuki, H., & Okubo, Y. (2014). Occupancy of serotonin transporter by tramadol: a positron emission tomography study with [11C]DASB. *The International Journal of Neuropsychopharmacology / Official Scientific Journal* *of the Collegium Internationale Neuropsychopharmacologicum (CINP)*, *17*(6), 845–50.

Ogden, R. T., Ojha, A., Erlandsson, K., Oquendo, M. A., Mann, J. J., and Parsey, R. V. (2007). I*n vivo* Quantification of Serotonin Transporters Using [^{11}C]DASB and Positron Emission Tomography in Humans: Modeling Considerations. Journal of Cerebral Blood Flow & Metabolism, 27(1):205–217.

Olesen, O. V., Sibomana, M., Keller, S. H., Andersen, F., Jensen, J., Holm, S., … Højgaard, L. (2009). Spatial resolution of the HRRT PET scanner using 3D-OSEM PSF reconstruction. *IEEE Nuclear Science Symposium Conference Record*, 3789–3790.

Parsey, R. V., Kent, J. M., Oquendo, M. A., et al (2006). Acute Occupancy of Brain Serotonin Transporter by Sertraline as Measured by [^{11}C]DASB and Positron Emission Tomography. Biological Psychiatry, 59(9):821–828.

Parsey RV, Ojha A, Ogden RT, Erlandsson K, Kumar D, Landgrebe M, Van Heertum R, Mann JJ. Metabolite considerations in the in vivo quantification of serotonin transporters using 11C-DASB and PET in humans. J Nucl Med. 2006 Nov;47(11):1796-802.

Parsey RV, Slifstein M, Hwang DR, Abi-Dargham A, Simpson N, Mawlawi O, Guo NN, Van Heertum R, Mann JJ, Laruelle M: Validation and reproducibility of measurement of 5-HT1A receptor parameters with [carbonyl-11C]WAY-100635 in humans: comparison of arterial and reference tissue input functions. J Cereb Blood Flow Metab 2000, 20(7):1111-1133.

Praschak-Rieder N1, Kennedy J, Wilson AA, Hussey D, Boovariwala A, Willeit M, Ginovart N, Tharmalingam S, Masellis M, Houle S, Meyer JH. Novel 5-HTTLPR allele associates with higher serotonin transporter binding in putamen: a [(11)C] DASB positron emission tomography study. Biol Psychiatry. 2007 Aug 15;62(4):327-31

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

Roussakis, A. A., Politis, M., Towey, D., and Piccini, P. (2016). Serotonin-to-dopamine transporter ratios in Parkinson disease. Neurology, 86(12):1152–1158.

Rylands AJ, Hinz R, Jones M, Holmes SE, Feldmann M, Brown G, McMahon AW, Talbot PS. Pre- and postsynaptic serotonergic differences in males with extreme levels of impulsive aggression without callous unemotional traits: a PET study using 11C-DASB and 11C-MDL100907. Biological Psychiatry 2012; 72:1004-1011.

Savli, M., Bauer, A., Mitterhauser, M., et al. (2012). Normative database of the serotonergic system in healthy subjects using multi-tracer PET. NeuroImage, 63(1):447–459.

Schwarz CG, Jones DT, Gunter JL, Lowe VJ, Vermuri PB, Senjem ML, Petersen RC, Knopman DS, Jack CR Jr; Alzheimer's Disease Neuroimaging Initiative. Contributions of imprecision in PET-MRI rigid registration to imprecision in amyloid PET SUVR measurements. Hum Brain Mapp. 2017 Apr 22. doi: 10.1002/hbm.23622. [Epub ahead of print]

Schain M, Varnäs K, Cselényi Z, Halldin C, Farde L, Varrone A. Evaluation of two automated methods for PET region of interest analysis. Neuroinformatics. 2014 Oct;12(4):551-62.

Shapiro, P. A., Sloan, R. P., Deochand, C., et al. (2014). Quantifying serotonin transporters by PET with [^{11}C]-DASB before and after interferon-alpha treatment. Synapse, 68(11):548–555.

Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, Frutiger S, Muley S, Laconte S, and Rottenberg D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage, 15, 747–71.

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

Sureau FC, Reader AJ, Comtat C, Leroy C, Ribeiro MJ, Buvat I, Trébossen R. Impact of image-space resolution modeling for studies with the high-resolution research tomograph. J Nucl Med, 2008; 49(6): 1000-8.

- Svarer C, Madsen K, Hasselbalch SG, Pinborg LH, Haugbøl S, Frøkjær VG, Holm S, Paulson OB, Knudsen GM. MR-based automatic delineation of volumes of interest in human brain PET-images using probability maps. NeuroImage 2005;24:969-79.

Tyrer, A. E., Levitan, R. D., Houle, S., Wilson, A. A., Nobrega, J. N., Rusjan, P. M., & Meyer, J. H. (2016). Serotonin transporter binding is reduced in seasonal affective disorder following light therapy. *Acta Psychiatrica Scandinavica*, *134*(5), 410–419.

van den Heuvel OA, Boellaard R, Veltman DJ, et al. Attenuation correction of PET activation studies in the presence of task-related motion. Neuroimage. 2003 Aug;19(4):1501-9.

- Whitley E & Ball J. (2002). Statistics review 4: Sample size calculations. Crit Care, 6(4);335-341.

Wu Y, Carson RE (2002) Noise reduction in the simplified reference tissue model for neuroreceptor functional imaging. J Cereb Blood Flow Metab 22:1440—1452.

Zanderigo, F., Mann, J. J., & Ogden, R. T. (2017). A hybrid deconvolution approach for estimation of in vivo non-displaceable binding for brain PET targets without a reference region. PLoS One. 2017 May 1;12(5):e0176636.

Zhang YW, Turk BE, Rudnick G. Control of serotonin transporter phosphorylation by conformational state. Proc Natl Acad Sci U S A. 2016 May 17;113(20):E2776-83.

APPENDIX C

Paper [C]

Nørgaard M, Greve DN, Svarer C, Strother SC, Knudsen GM, Ganz M. The Impact of Preprocessing Pipeline Choice in Univariate and Multivariate Analyses of PET Data. *Pattern Recognition in Neuroimaging* (PRNI), IEEE Explore, 2018, pp. 1-4. DOI: 10.1109/PRNI.2018.8423962

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 ± 30 (mean ± 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 ± 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 ($\mathrm{BP_{ND}}$) 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 $\mathrm{BP_{ND}}$'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) $\mathrm{BP_{ND}}$. 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:

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

where $\bar{\mathbf{X}}^k_{\mathbf{train}}$ is the training data mean from class C_k, and $\mathbf{L}_{\mathbf{train}}$ 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.

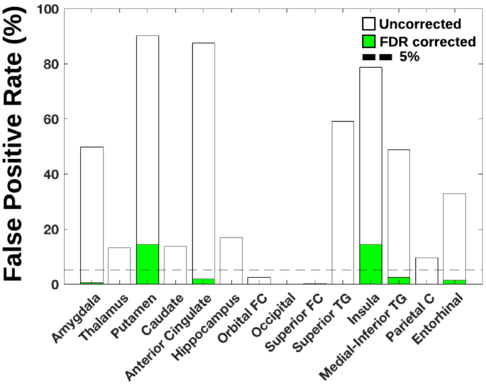

Fig. 1. Number of significant results (paired t-test, $P < 0.05$) in 384 pipelines divided by 384, expressed as a percentage for 14 brain regions. Blank is not corrected for multiple comparisons, whereas green is corrected using FDR.

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

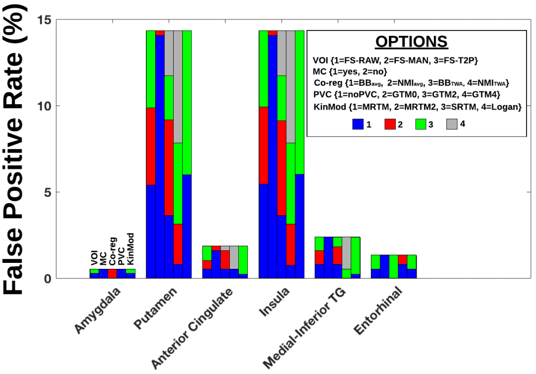

Fig. 2. Number of significant results (paired t-test, $P < 0.05$) in 384 pipelines divided by 384, expressed as a percentage for 14 brain regions (corrected for multiple comparisons at FDR=0.05 within each pipeline). The five vertical bars within each region represent the distribution of choices, and 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).

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

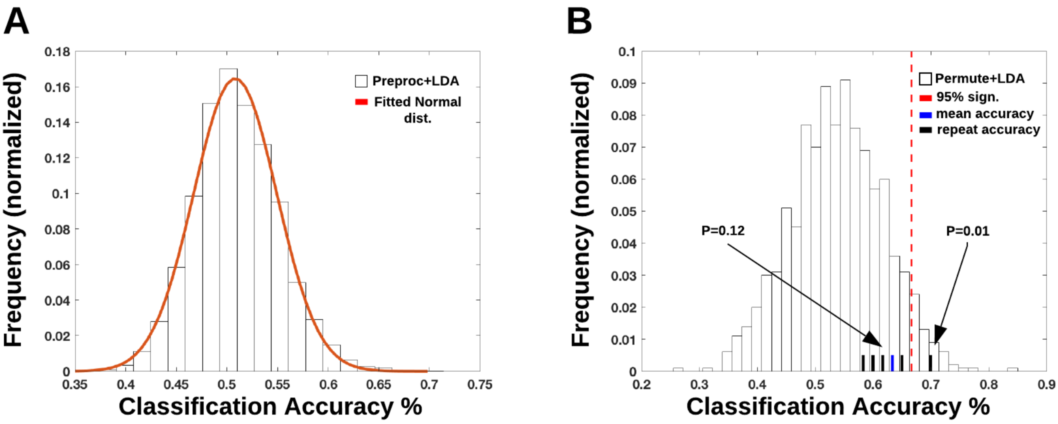

Fig. 3. **(A)** Normalized distribution of classification accuracies (%) for 10 times repeated 5-fold cross-validation and for 384 different preprocessing choices **(B)** Normalized distribution of 1000 permuted classification accuracies (%) for the pipeline maximizing the classification accuracy in (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.

Appendix D

Paper [D]

Nørgaard M, Ganz M, Svarer C, Douglas N. Greve, Vibe G. Frokjaer, Strother SC, Knudsen GM. The Impact of Different Preprocessing Strategies in PET Neuroimaging: A [^{11}C]DASB-PET Study. Submitted to *Journal of Cerebral Blood Flow and Metabolism.*

The Impact of Different Preprocessing Strategies in PET Neuroimaging: A [^{11}C]DASB-PET Case

Running Title: Different Preprocessing Strategies in PET Neuroimaging Lead to Different Conclusions

Martin Nørgaard[1,2]
Melanie Ganz[1,3]
Claus Svarer[1]
Vibe G. Frokjaer[1]
Douglas N. Greve[5]
Stephen C. Strother[4]
Gitte M. Knudsen[1,2*]

[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, Baycrest, and Department of Medical Biophysics, University of Toronto, Toronto, Canada
[5] Athinoula A. Martinos Center for Biomedical Imaging,
Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

* Corresponding author gmk@nru.dk

Abstract

Positron Emission Tomography (PET) neuroimaging provides unique possibilities to study biological processes *in vivo* under basal and interventional conditions. For quantification of PET data, researchers apply different arrays of sequential data analytic methods ("preprocessing strategy", also referred to as a "pipeline"), but it is unknown how the choice of preprocessing strategy affects the final outcome.

Here, we use an available data set from a double-blind, randomized, placebo-controlled [^{11}C]DASB-PET study as a case to evaluate how the choice of preprocessing strategy affects the outcome of the study. We test the impact of 384 commonly used preprocessing strategies on a previously reported positive association between the change from baseline in neocortical serotonin transporter binding determined with [^{11}C]DASB-PET, and change in depressive symptoms, following a pharmacological sex hormone manipulation intervention in 30 women.

We find that 36% of our preprocessing strategies replicate the originally reported finding ($p < 0.05$), meaning that 64% of preprocessing strategies do not result in a statistically significant association. The two preprocessing steps that were most critical for the outcome were motion correction and kinetic modeling of the dynamic PET data.

In conclusion, the choice of preprocessing strategy can have a major impact on a study outcome.

Key words: Positron Emission Tomography; preprocessing; head motion; partial volume correction; kinetic modeling; pharmacological intervention; [^{11}C]DASB

INTRODUCTION

Science is entering a reproducibility crisis (Baker 2016). Historically this has meant being unable to reproduce scientific results in an independent sample, even when using the same experimental design and methodological choices (Open Science Collaboration 2015).
In practice, the outcome of two similar studies are never 100% overlapping because of differences in methodology, e.g., available equipment, settings, and sample data (Goodman et al. 2016).
Apart from this, it is also challenging to identify the sources of variation that originate from each methodological choice, and how it may ultimately influence the study outcome. Arriving at a plausible conclusion is, often wrongly, taken as justification of the methodological choices made, providing a systematic bias toward prevailing scientific expectations (Strother et al. 2002).

In Positron Emission Tomography (PET) neuroscience, only a few studies have investigated the impact of methodological choices on the outcome of a study. Samper-González and coworkers (2018) assessed if the preprocessing strategy of FDG-PET data affected the classification of patients suspected of Alzheimers Disease, and found no differences in predictive performance when switching preprocessing strategy to, e.g., a new atlas, different levels of spatial smoothing, or application of partial volume correction (PVC). In contrast, Greve et al. 2016 showed that different PVC methods led to different conclusions, and that extreme care should be taken when applying PVC. The effect of PVC has also been documented by previous studies (Berkouk et al. 1996 and Meltzer et al. 1996).

Mukherjee et al. (2016) investigated the effects of frame-based correction of head motion in PET brain imaging, and showed that head motion can cause significant degradation of the image quality. The argument that head motion in PET brain imaging renders PET data disturbed or even useless has been made before (Olesen et al. 2013, Anton-Rodriguez et al. 2010). More recently, Nørgaard et al. (2018a) showed in a meta-analysis including 105 publications that between-subject variability of striatal serotonin transporter (5-HTT) binding, as imaged with [^{11}C]DASB-PET, was lower when motion correction (MC) was done and that it translated into 26% fewer subjects needed in a group analysis to achieve similarly powered statistical tests. In spite of these observations, many recent studies do not include MC in their preprocessing strategy (e.g., Kim et al. 2016, Zientek et al. 2016, Hinderberger et al. 2016, Frick et al. 2016).

Recently, we showed that inconsistent reports of 5-HTT levels in healthy individuals might be explained by variations in acquisition and preprocessing strategy (Nørgaard et al. 2018a).

However, while it may be inevitable that different methods are applied in different PET centres, the key question is how these differences affect the outcome of a study?

Here, we investigate how the outcome depends on the choice of preprocessing strategy. We use data from Frokjaer et al. 2015 which is a double blind, randomized, placebo-controlled intervention study of 60 healthy women. We applied 384 different preprocessing strategies to test which of them reproduce the main outcome from Frokjaer et al. 2015, namely a positive association between the emergence of depressive symptoms and change in cerebral 5-HTT binding following a pharmacological se-hormone manipulation with a gonadotropin-releasing hormone agonist (GnRHa) intervention. In addition, we also tested how preprocessing strategy would influence the association between the personality trait neuroticism and change in 5-HTT binding from baseline, which was also part of the original analysis (Frokjaer et al. 2015). Because preprocessing strategies in the [^{11}C]DASB-PET literature have been assumed to produce near similar results (Kim et al. 2006, Ginovart et al. 2001, Ogden et al. 2006), we hypothesized that by across a range of (reasonable) preprocessing strategies, the study conclusions would remain the same (i.e. the conclusions are preprocessing independent).

METHODS

1.1 Participants

A total of 60 female participants (mean age 24.3 ± 4.9 years) were included in a double-blind, randomized, placebo-controlled study (Frokjaer et al. 2015), which investigated depressive responses to sex-steroid hormone manipulation and related brain imaging signatures. Participants received either a subcutaneouos injection of a gonadotropin releasing hormone agonist (GnRHa) implant (ZOLADEX with 3.6 mg of goserelin; Astra Zeneca, London, UK) (N=30) or saline (N=30). We provide demographic information in the supplementary (Table S1). One subject in the GnRHa group was excluded due to an issue with the PET acquisition, leaving 29 subjects available for analysis. Further details 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) and registered as a clinical trial: www.clinicaltrials.gov under the trial ID NCT02661789. All participants gave written informed consent.

1.2 Positron Emission Tomography

All participants were scanned in a Siemens ECAT HRRT scanner with the selective 5-HTT radioligand [^{11}C]DASB (Houle et al. 2000). The protocol consisted of a 90 minutes dynamic acquisition (3D list-mode) post injection of 587±30 (mean ± SD) MBq bolus into an elbow vein. The PET data was reconstructed into 36 frames (6x10, 3x20, 6x40, 5x60, 5x120, 8x300, 3x600 seconds) using a 3D-OSEM-PSF algorithm with TXTV attenuation correction (Sureau et al. 2008, Keller et al. 2013).

Reconstructed dynamic PET images contain the concentration of radioactivity (Bq/mL) as a function of time (time-activity curve, TAC) from each voxel or brain region.

1.3 Magnetic Resonance Imaging

An isotropic T1-weigthed MP-RAGE was acquired for all participants (matrix size = 256 x 256 x 192; voxel size = 1 mm; TR/TE/TI = 1550/3.04/800 ms; flip angle = 9°) using either a Siemens Magnetom Trio 3T or a Siemens 3T Verio MR scanner. Furthermore, an isotropic T2-weighted sequence (matrix size 256 x 256 x 176; voxel size = 1 mm; TR/TE = 3200/409 ms; flip angle = 120˚) was acquired for all participants. All acquired MRI's were corrected for gradient nonlinearities (Jovicich et al. 2006), and examined to ensure absence of structural abnormalities.

1.4 Preprocessing steps for PET and MRI

Brain 5-HTT binding was estimated by applying a preprocessing strategy consisting of a fixed sequence of five steps (MC, co-registration, delineation of volumes of interest (VOI), Partial Volume Correction (PVC) and kinetic modeling) with each step consisting of 2-4 choices.

All preprocessing strategies have previously been applied and evaulated (Nørgaard et al. 2018b). The steps are listed below in the order in which they were applied, producing a total of 384 different preprocessing strategies. The outcome measure for each preprocessing strategy is an estimate of the brain regional non-displaceable binding potential (BP_{ND}) (Innis et al. 2007).

Further details on all preprocessing steps can be found in Nørgaard et al. 2018b.

1.4.1 Motion Correction

The PET data was analyzed either with or without MC (nMC). The MC was carried out using AIR (v. 5.2.5). First, alignment parameters for PET frame 10-36 to a frame with high signal-to-noise ratio (frame 26) were estimated and secondly, each frame was resliced into a motion corrected 4D data set (Frokjaer et al. 2015).

1.4.2 Co-Registration

All single-subject 4D PET images were either summed or averaged across frames to estimate either a time-weighted (twa) or averaged over all frames (avg) 3D image for co-registration. The two co-registration techniques Normalized Mutual Information (NMI, Studholme et al. 1999) or Boundary-Based Registration (BBR, Greve et al. 2009) were subsequently applied to either the twa or the avg image.

All MRI's were co-registered to native PET space for subsequent analysis.

1.4.3 Delineation of Volumes of Interest

All MRI's were processed (recon-all) using FreeSurfer (http://surfer.nmr.mgh.harvard.edu, version 5.3) (Fischl et al. 2004). After running the FreeSurfer (FS) pipeline, manual edits can be applied to correct for errors in the delineation. In addition, if a T2-weighted image is available, the FS pipeline can be re-run with T2-optimization for removal of errors in the delineation of regions. All three choices of FS processing were carried out, and we refer to these as FS-RAW (standard output), FS-MAN (output with manual edits) and FS-T2P (output with T2-optimization). The VOI's neocortex, anterior cingulate cortex (ACC), striatum and midbrain were used for comparison with Frokjaer et al. 2015. The neocortex region was generated by taking all cortical TACs in the Desikan-Kiliany atlas provided by FreeSurfer (total of N = 68 regions across both hemispheres) and volume-weighting them into a single neocortical TAC. This can be expressed as

$$TAC_{neocortex} = \frac{\sum_{i=1}^{N} TAC_i \times volume_i}{volume_{total}}$$

The striatum was generated by averaging the regions putamen and caudate (Tuominen et al. 2017). The remaining regions, ACC and midbrain, were automatically generated by FS.

1.4.4 Partial Volume Correction

The PET data was corrected either without (noPVC) or with PVC. The VOI-based PVC technique, Geometric Transfer Matrix (GTM) by Rousset et al. 1998, was applied using PETsurfer (surfer.nmr.mgh.harvard.edu/fswiki/PetSurfer, Greve et al. 2016) using three different assumptions of the point spread function (PSF) of the PET scanner.

Because the PSF for a HRRT scanner varies depending on the distance from the center of field-of-view (Olesen et al. 2009), the application of PVC was carried out using the PSF settings: 0 mm, 2 mm or 4 mm. This results in four strategies for the PVC preprocessing step.

1.4.5 Kinetic Modeling

Four kinetic models were applied, all based on reference tissue modeling (RTM) and implemented in MATLAB 2016b (https://www.mathworks.com) for parallel execution purposes. All models used cerebellum (excluding vermis) as a reference region. The Multilinear Reference Tissue Model (MRTM) and Multilinear Reference Tissue Model 2 (MRTM2) were applied as described in Ichise et al. 2003. The non-invasive Logan reference tissue model was applied as described in Logan et al. 1996. The Simplified Reference Tissue Model (SRTM) was applied as described in Lammertsma and Hume 1996. All implemented models were validated with PMOD v. 3.0 (10 subjects < 0.1% difference in BP_{ND}).

1.5 Statistics

Linear regression models were applied with BP_{ND} as the independent variable (separate models for each region) and either neuroticism score or Hamiltons Depression score as the dependent variable. This sums to 4 regions x 2 dependent variables x 384 preprocessing strategies = 3072 linear regression models. All analyses were performed in MATLAB 2016b (www.mathworks.com). P-values below .05 were considered statistically significant.

RESULTS

Regional analysis of BP_{ND} and across preprocessing strategies

Table 1 summarizes the regional group mean BP_{ND} results across 384 preprocessing strategies, and provides a statistical comparison (two sample t-tests) at baseline between the placebo and GnRHa group. The percentage of preprocessing strategies resulting in $p < 0.05$ is the number of instances out of 384 preprocessing strategies where we identified a significant difference between groups ($p < 0.05$) at baseline.

	Placebo (n = 30)	**GnRHa (n = 29)**	**GnRHa versus placebo p-value**	**% preprocessing strategies resulting in p < 0.05**

Neocortex	0.98±.46	0.95±.46	0.25±.12	0
ACC	1.39±.46	1.32±.46	0.17±.60	0.5
Striatum	2.69±.34	2.51±.35	0.21±.12	11.5
Midbrain	2.27±.34	2.24±.36	0.73±.18	0

Table 1: [^{11}C]DASB BP_{ND} in different brain regions in placebo versus active treatment at baseline. Regional BP_{ND}'s are given as mean ± SD resulting from 384 preprocessing strategies. ACC: anterior cingulate.

Depressive Symptoms and change in [^{11}C]DASB Binding from baseline Across Preprocessing Strategies

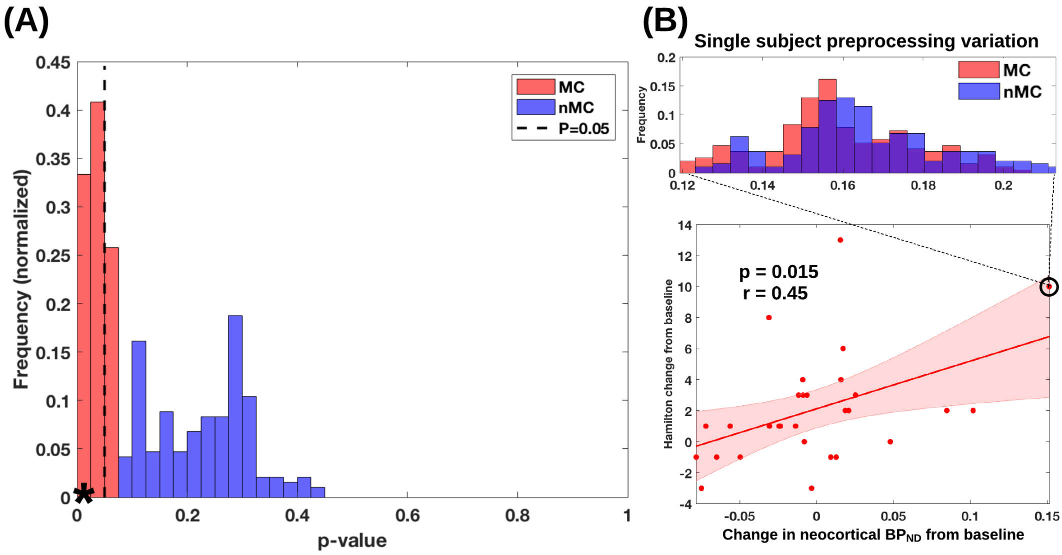

Figure 1: (A) Histogram of p-values obtained across 384 preprocessing strategies examining the association between change in neocortical BP_{ND} and change in Hamilton score from baseline in the GnRHa group. MC = 'Motion Correction', nMC = 'no Motion Correction', SRTM = 'Simplified Reference Tissue Model' (B) Lower plot shows the association between the change in neocortical BP_{ND} and Hamilton score from baseline ($p = 0.015$), using the recommended preprocessing strategy from Nørgaard et al. 2018b (black star in (A)). The shaded error bar (B, lower) indicates the 95% confidence interval of the starred result (inferential bounds). Of the 384 preprocessing strategies, 36% were significant at $p < 0.05$ and they all included MC. The black circle (B, lower) and the histogram (B, upper) illustrate the variation (between 0.12 and 0.22) in the change in neocortical BP_{ND} from baseline for a single subject, across the 384 preprocessing strategies.

Figure 1 shows a summary of the main results of this study. Figure 1A shows the frequency of preprocessing strategies as a function of p-value for the association between Hamilton change from baseline and change in neocortical BP_{ND} from baseline. The vertical dashed black line is the cut-off for $p < 0.05$. Effect sizes (i.e. Pearson's correlation) varied from 0.15 to 0.45 (Figure S1). Figure 1B (lower plot) shows the association for a single preprocessing strategy highlighted by the black star in Figure 1A. The shaded error bars represent the 95% confidence interval (inferential bounds). The upper plot in Figure 1B shows how the change in BP_{ND} from baseline varies as function of preprocessing strategy for a single subject, as marked by the black circle.

The remaining histograms of p-values as a function of preprocessing strategy can be found in the supplementary material.

Neuroticism and 5-HTT Binding Across Preprocessing Strategies

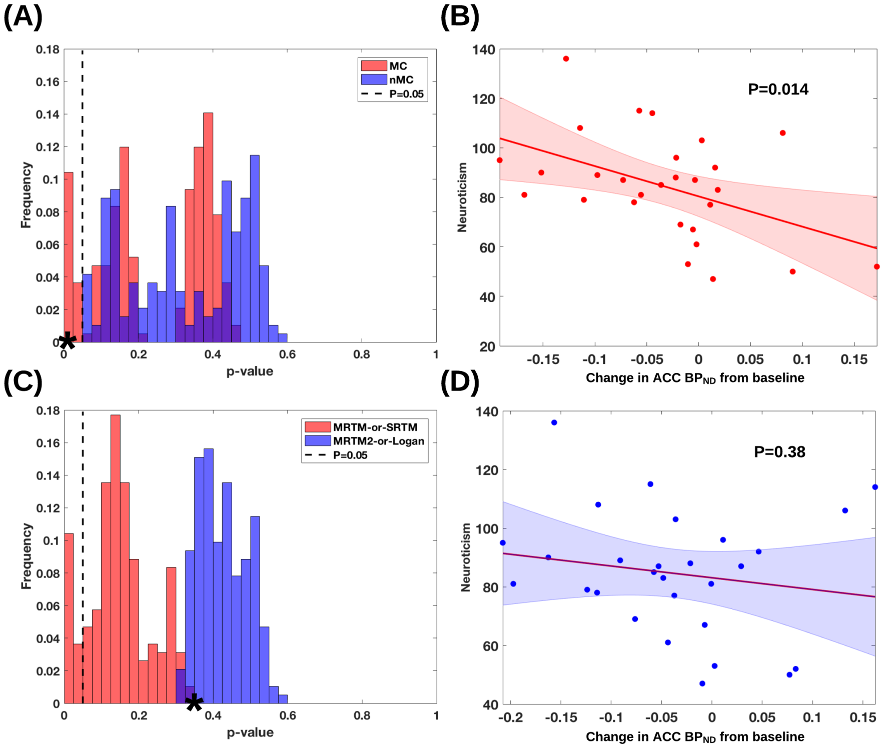

Figure 2: (A) Histogram of obtained p-values for the association between the change in ACC BP_{ND} from baseline and neuroticism, in the GnRHa group and across 384 preprocessing strategies. MC =

‘Motion Correction’, nMC = ‘no Motion Correction’. **(B)** association between the increase in ACC BP_{ND} from baseline and neuroticism ($p = 0.014$), using one of the 27 preprocessing strategies (black star in (A)) yielding a significant correlation ($p < 0.05$). All preprocessing strategies yielding statistically significant outcomes share the steps MC and SRTM. **(C)** Similar histogram as in (A) but now divided into SRTM-or-MRTM (red) and MRTM2-or-Logan (blue) **(D)** similar plot as in (B) but for a pipeline that generates a statistically non-significant outcome (black star in (C)). Abbreviations: MC=’Motion Correction’, SRTM=’Simplified Reference Tissue Model’, MRTM=’Multilinear Reference Tissue Model’, ACC=’Anterior Cingulate Cortex’.

Figure 2A shows the frequency of preprocessing strategies as a function of p-value for the association between neuroticism and change in ACC BP_{ND} from baseline. The strategies are split into those with MC (red) and those without MC (blue). The vertical dashed black line is the cut-off for $p < 0.05$. Figure 2B shows the association between neuroticism at baseline and change in ACC BP_{ND} from baseline for a single preprocessing strategy generating a p-value of 0.014. Figure 2C shows the effects of kinetic modeling choice on the frequency of p-values for the association between neuroticism and change in ACC BP_{ND} from baseline. The red distribution is for choices of MRTM and SRTM, whereas the blue distribution is for choices of MRTM2 and non-invasive Logan. Figure 2D shows the latter association for a single preprocessing strategy as marked by the black star in Figure 2C ($p = 0.38$).

The supplementary material contains p-values from all 3072 linear regression models in freely available MATLAB files (*.mat).

DISCUSSION

The present analysis is to our knowledge the first to systematically examine the effects of several preprocessing interactions on the outcome of an *in vivo* PET neuroimaging study. Our study builds on data regarding behavioural phenotypes and cerebral 5-HTT and we find that different preprocessing strategies result in different outcomes when it comes to the emergence of depressive symptoms and changes in cerebral 5-HTT after a sex hormone intervention.

Consistent with previous studies (e.g., Olesen et al. 2013, Anton-Rodriguez et al. 2010), we identify MC as the main key step in the preprocessing strategy. It has been estimated that motion artefacts are present in 10-20% of high-resolution PET data (Ooi et al. 2009), and noise confounds are amplified during long acquisition scans (van der Kouwe et al. 2006). Nørgaard et al. 2018a showed that 40% of all published [^{11}C]DASB-PET studies left out MC in their preprocessing strategy, and that MC resulted in a reduced between-subject variability compared to data without MC. Our finding is particularly interesting because the PET data we used were carefully selected to ensure minimal head motion (< 3 mm median movement).

In a recent study, we outlined methodological differences in healthy individuals with test-retest and estimated their impact on performance metrics of bias, within- and between-subject variability (Nørgaard et al. 2018b). Based on the estimated variabilities, we provided recommendations on the optimal preprocessing strategy, so maximally powered results could be obtained depending on the study design. When the recommendations in Nørgaard et al. 2018b are followed in the analysis of the present independent data set, we arrive at the same conclusions as made in the original paper by Frokjaer et al. 2015.

The same study also showed that the preprocessing steps MC, PVC and kinetic modeling were the most prominent components that contribute to the level of within- and between-subject variance. In the present study, we replicate that MC (Figure 1) and kinetic modeling (Figure 2) have profound effects on the results. Notably, the combination of nMC and SRTM-or-MRTM eliminated the significant correlation between neuroticism and 5-HTT levels in the ACC (Figure 2). Two previous [^{11}C]DASB-PET studies combined nMC and SRTM in their analysis (Nogami et al. 2013, Ogawa et al. 2014); the remaining 4 studies applied MC before SRTM (Comley et al. 2013, Turkheimer et al. 2012, Abanades et al. 2011, Hammoud et al. 2010). Needless to say, we do not know what the outcome of the two first studies would have been, had MC been done.

SRTM uses non-linear least squares optimization to estimate the model parameters, and it is likely that when combined with nMC this may result in an unstable local-maxima solution due to increased noise.

Another notable observation was that the single-subject variability resulting from preprocessing strategy was nearly as large as the between-subject variability (Figure 1B, upper). Under the assumption that the majority of preprocessing strategies are equally valid (or used), this suggests that single subject variability across preprocessing choices should be taken into account when interpreting the robustness of the observed associations. This will be particularly critical in studies where it can be expected that a smaller (sensitive) subgroup of the population drives the observed association as is the case in the present example; a subgroup of women appeared to be particularly sensitive to sex-hormone manipulation whereas the majority of women balanced the intervention quite well in terms of developing depressive symptoms.

We also tested how preprocessing strategy would influence the statistical significance of the association between the personality trait neuroticism and change in 5-HTT binding from baseline and the potential dependency on intervention, which was also part of the original analysis (Frokjaer et al. 2015). We found that 27 out of 384 preprocessing strategies resulted in a statistically significant negative correlation between neuroticism and change in ACC 5-HTT from baseline in the intervention group (Figure 2). While neuroticism has consistently been implicated in stress regulation, depression and brain 5-HTT (Tuominen et al. 2017, Hirvonen et al. 2015), there may also be some aspects of neuroticism as a trait that potentially could affect the cerebral 5-HTT levels when PET-scanned twice.

Based on previous studies, the serotonin system and stress regulation system appear to be intimately related (Frokjaer et al. 2013, Frokjaer et al. 2014, Jacobsen et al. 2016). In general, acute stress enhances serotonin output, and in turn, serotonin signaling influences the secretion of corticosteroids (Lanfumey et al. 2008, Kim et al. 2006). Assuming/speculating that it may be less stressful to participate in a PET scan for the second time, an index of stress coping capacity, as neuroticism, should matter in terms of baseline to follow-up differences in 5-HTT binding. This may offer an explanation for why we and others found that in the absence of any interventions, the cerebral 5-HTT was lower when healthy volunteers were scanned the second time relative to baseline (Nørgaard et al. 2018b, Kim et al. 2006). To test this hypothesis, we carried out a post-hoc exploratory analysis investigating whether we could find a group interaction effect between neuroticism and change in BP_{ND}. The expected interaction effect was found (Figure S1 in the supplementary) for some but not all regions and preprocessing strategies. The regions included the amygdala, putamen, ACC and superior temporal gyrus, and the association was mainly driven by preprocessing strategies containing MC and SRTM/MRTM (all results provided in the supplementary). The results suggest that the particular GnRH intervention disrupts the expected neruoticism dependent variation between baseline and 5-HTT binding and is in line with other observations (Stenbæk et al. 2019). However, it was clearly not the scope of this article to further address the potential mechanistics of this phenomenon. We also considered if the first scan sessions, i.e. expected higher stress levels, would be associated with more head motion, but we did not find any differences in motion between the two scan sessions across intervention groups (data not shown). Further studies should elucidate if perceived stress or indices of stress sensitivity can explain test-retest effects in longitudinal PET studies and if such observations translate to other markers of serotonin signaling.

While we highlight in this study that different preprocessing strategies give rise to different outcomes, there are also some statistical considerations that could help neuroscientists to mitigate towards more predictive and replicable science. In the current data set, a more predictive and reproducible analysis would have been obtained by the application of a predictive model evaluated in a cross-validation framework instead of applying a correlational analysis. Predictive models that provide a predictive accuracy are conceptually intriguing as they provide a measure of the ability to correctly predict the experimental condition and/or behaviour in an independent sample. In our case, a correlational analysis corresponds to a fixed effect or association model, and the outcome can only be interpreted with respect to the given data set (Gabrieli et al. 2015). In contrast, a predictive analysis using cross-validation corresponds to identifying the associations that can generalize to the population (i.e. random effect model). Nevertheless, a plausible explanation using a correlational analysis is often chosen over predictive accuracy, but may have limited ability to generalize to an independent sample (Yarkoni and Westfall 2017).

To further increase generalizability of an outcome, the current preprocessing framework could also be used to estimate the expected outcome conditioned over multiple preprocessing strategies (i.e. have 36% confidence in the outcome). The estimated expectation will provide a confidence in the extent to which the generated outcome is valid across preprocessing strategies. The expected outcome conditioned over preprocessing strategies should help to control the probability that the outcome could arise under the null hypothesis (false discovery rate), but it does not necessarily impose the generally (and abitrarily) required probability be less than 5% for publication (Greve et al. 2017, Benjamin et al. 2017). Just to make it clear: We do not propose that in all future PET studies, researchers should test a full range of preprocessing strategies before concluding on the outcome. We will, however, emphasize that it is recommended to verify that an outcome is not driven by the result of a single preprocessing strategy.

From a statistical standpoint, the expected outcome conditioned over preprocessing strategies is not sufficient to correct for the number of tested preprocessing strategies, nor does it answer whether preprocessing strategies are significantly different from each other. Developing such a statistical framework including a predictive component would be of great value for the neuroimaging community, but is currently considered as future work.

Our study is not without limitations. First, the subset of 384 preprocessing strategies of all possible preprocessing strategies, does not allow us to infer whether the expected outcome conditioned over preprocessing strategies may be either negatively or positively biased. As shown by Nørgaard et al. 2018a, there exist at least 21,150,720 PET neuroimaging workflows (data acquisition and preprocessing), so it is not unlikely that the current sampling distribution for the expected outcome conditioned over preprocessing strategies does not fully represent the true underlying distribution. Another limitation in the study is that all the different choices is tested using one single framework, for the effect of MC using the AIR 5.3.0 package (Woods et al. 1992) and for other processing tools using FreeSurfer (http://surfer.nmr.mgh.harvard.edu). There is of course many other possibilities for using other packages for these steps which potentially could lead to other results. We note, however, that this dillemma currently holds true in all fields of neuroimaging, and for scientific workflows in general, that have highly varying methodology being applied with limited ability to reproduce previous findings, especially in studies with low sample sizes.

CONCLUSION

In conclusion, we find that different preprocessing strategies lead to different conclusions, which illustrates that it is important to consider and to declare preprocessing strategies before analyzing the data. Even in the absence of larger head movements within the scanner, MC and kinetic modeling of dynamic PET data seem to be the most important steps. Future studies are needed to explicitly rule out potential external variables related to data acquisition and/or preprocessing that may govern the outcome of a study.

ACKNOWLEDGEMENTS

We wish to thank all the participants for kindly joining the research project. We thank the John and Birthe Meyer Foundation for the donation of the cyclotron and PET scanner. Peter Steen Jensen and Vincent Beliveau are gratefully acknowledged. MN was supported by the National Institutes of Health (Grant 5R21EB018964-02), the Lundbeck Foundation (Grant R90-A7722), and the Independent Research Fund Denmark (DFF-1331-00109 & DFF-4183-00627).

REFERENCES

Anton-Rodriguez JM, M. Sibomana, M.D. Walker, M.C. Huisman, J.C. Matthews, M. Feldmann, S.H. Keller, and M.C. Asselin. Investigation of motion induced errors in scatter correction for the HRRT brain scanner. IEEE Nuclear Science Symposium Conference Record, pages 2935–2940, 2010.

Baker M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452-454, 2016.

Benjamin, D.J., Berger, J.O., Johnson, V.E., Sep 1, 2017. Commentary: redefine statistical significance. Nature Human Behaviour.

Berkouk, K., Quarantelli, M., Prinster, A., Landeau, B., Alfano, B., & Baron, J. C. (2006). Mapping the relative contribution of gray matter activity vs. volume in brain PET: A new approach. Journal of Neuroimaging, 16(3), 224–235.

Fanelli D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proceedings of the National Academy of Sciences (PNAS), 2018;115(11): 2628-2631.

Fischl B. FreeSurfer. Neuroimage. 2012 Aug 15; 62(2): 774-781.

Frick, A., Åhs, F., Engman, J., Jonasson, M., Alaie, I., Björkstrand, J., Frans, Ö., Faria, V., Linnman, C., Appel, L., Wahlstedt, K., Lubberink, M., Fredrikson, M., and Furmark, T. (2015). Serotonin Synthesis and Reuptake in Social Anxiety Disorder. JAMA Psychiatry, (JUNE):E1–E9.

Frokjaer, V. G., Erritzoe, D., Holst, K. K., Jensen, P. S., Rasmussen, P. M., Fisher, P. M., … Knudsen, G. M. (2013). Prefrontal serotonin transporter availability is positively associated with the cortisol awakening response. European Neuropsychopharmacology, 23(4), 285–294

Frokjaer, V. G., Erritzoe, D., Holst, K. K., Madsen, K. S., Fisher, P. M., Madsen, J., … Knudsen, G. M. (2014). In abstinent MDMA users the cortisol awakening response is off-set but associated with prefrontal serotonin transporter binding as in non-users. The International Journal of Neuropsychopharmacology, 17(08), 1119–1128.

Frokjaer, V. G., Pinborg, A., Holst, K. K., Overgaard, A., Henningsson, S., Heede, M., Larsen, E. C., Jensen, P. S., Agn, M., Nielsen, A. P., Stenbæk, D. S., Da Cunha-Bang, S., Lehel, S., Siebner, H. R., Mikkelsen, J. D., Svarer, C., and Knudsen, G. M. (2015). Role of serotonin transporter changes in depressive responses to sex-steroid hormone manipulation: A positron emission tomography study. Biological Psychiatry, 78(8):534–543.

Gabrieli, J. D. E., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron, 85(1), 11–26.

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, 21(11):1342–1353.

Goodman SN, Fanelli D, Ioannidis JPA (2016). What does research reproducibility mean? Sci Transl Med 8:341ps312.

Green MV, J. Seidel, S.D. Stein, T.E. Tedder, K.M. Kempner, C. Kertzman, and T.A. Zeffiro. Head movement in normal subjects during simulated PET brain imaging with and without head restraint. Journal of Nuclear Medicine, 35(9):1538–1546, 1994.

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

Greve, D. N., Salat, D. H., Bowen, S. L., Izquierdo-Garcia, D., Schultz, A. P., Catana, C., Becker, J. A., Svarer, C., Knudsen, G. M., Sperling, R. A., and Johnson, K. A. (2016). Different partial volume correction methods lead to different conclusions: An 18F-FDG-PET study of aging. NeuroImage, 132:334–343.

Ichise, M., Liow, J.-S., Lu, J.-Q., Takano, A., Model, K., Toyama, H., … Carson, R. E. (2003). Linearized reference tissue parametric imaging methods: application to [11C]DASB positron emission tomography studies of the serotonin transporter in human brain. *Journal of Cerebral Blood Flow and Metabolism*, *23*(9), 1096–1112.

Jakobsen, G. R., Fisher, P. M., Dyssegaard, A., McMahon, B., Holst, K. K., Lehel, S., … Frokjaer, V. G. (2016). Brain serotonin 4 receptor binding is associated with the cortisol awakening response. Psychoneuroendocrinology, 67, 124–132.

Jovicich J, Czanner S, Greve DN, Haley E, van der Kouwe A, Gollub R, Kennedy D, et al. Reliability in Multi–Site Structural MRI Studies: Effects of Gradient Non–Linearity Correction on Phantom and Human Data. NeuroImage, 2006; 30(2): 436–43.

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.

Kim, J. S., Ichise, M., Sangare, J., and Innis, R. B. (2006). PET Imaging of Serotonin Transporters with [11C]DASB: Test- Retest Reproducibility Using a Multilinear Reference Tissue Parametric Imaging Method. J. Nucl. Med., 47(2):208–214.

Lanfumey, L., Mongeau, R., Cohen-Salmon, C., & Hamon, M. (2008). Corticosteroid-serotonin interactions in the neurobiological mechanisms of stress-related disorders. Neuroscience and Biobehavioral Reviews, 32(6), 1174–1184.

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

Logan J, Fowler JS, Volkow ND, Wang GJ, Ding YS, Alexoff DL: Distribution volume ratios without blood sampling from graphical analysis of PET data. J Cereb Blood Flow Metab 1996, 16(5):834-840.

Meltzer, C. C., Zubieta, J., Brandt, J., Tune, L., Mayberg, H., & Frost, J. (1996). Regional hypometabolism in Alzheimer's disease as measured by PET following correction for effects of partial volume averaging. Neurology, 47, 454–461.

Nørgaard M, Ganz M, Svarer C, Feng L, Ichise M, Lanzenberger R, Lubberink M, Parsey RV, Politis M, Rabiner EA, Slifstein M, Sossi V, Suhara T, Talbot PS, Turkheimer F, Strother SC, Knudsen GM. 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. Epub ahead of print.

Nørgaard M, Ganz M, Svarer C, Frokjaer VG, Greve DN, Strother SC, Knudsen GM. Optimization of Preprocessing Strategies in Positron Emission Tomography (PET) Neuroimaging: A [^{11}C]DASB Study. In revision.

Ogden, R. T., Ojha, A., Erlandsson, K., Oquendo, M. A., Mann, J. J., and Parsey, R. V. (2007). I*n vivo* Quantification of Serotonin Transporters Using [^{11}C]DASB and Positron Emission Tomography in Humans: Modeling Considerations. Journal of Cerebral Blood Flow & Metabolism, 27(1):205–217.

Olesen OV, Sullivan JM, Mulnix T, Paulsen RR, Højgaard L, Roed B, Carson RE, Morris ED, Larsen R. List-mode PET motion correction using markerless head tracking: proof-of-concept with scans of human subject. IEEE Trans Med Imaging. 2013 Feb;32(2):200-9.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science 349:aac4716.

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

Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, Bacci M, Wen J, Bertrand A, Bertin H, Habert MO, Durrleman S, Evgeniou T, Colliot O; Alzheimer's Disease Neuroimaging Initiative; Australian Imaging Biomarkers and Lifestyle flagship study of ageing. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. Neuroimage. 2018 Dec;183:504-521

Stenbæk, D. S., Budtz-Jørgensen, E., Pinborg, A., Jensen, P. S., & Frokjaer, V. G. (2019). Neuroticism modulates mood responses to pharmacological sex hormone manipulation in healthy women. Psychoneuroendocrinology, 99(October 2018), 251–256.

Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, Frutiger S, Muley S, Laconte S, and Rottenberg D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage, 15, 747–71.

Studholme C, Hill DLG, and Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn. 32, pp. 71–86, 1999.

Sureau FC, Reader AJ, Comtat C, Leroy C, Ribeiro MJ, Buvat I, Trébossen R. Impact of image-space resolution modeling for studies with the high-resolution research tomograph. J Nucl Med, 2008; 49(6): 1000-8.

Van Der Kouwe, A. J. W., Benner, T., & Dale, A. M. (2006). Real-time rigid body motion correction and shimming using cloverleaf navigators. Magnetic Resonance in Medicine, 56(5), 1019–1032.

Woods RP, Cherry SR, J. C. M. (1992). Rapid Automated Algorithm for Aligning and Reslicing PET Images. Journal of Computer Assisted Tomography, 620–33.

Appendix E

Paper [E]

Nørgaard M, Ozenne B, Svarer C, Frokjaer V, Ganz M. Preprocessing, Prediction and Significance: Framework and Application to Brain Imaging. Submitted to Medical Image Computing and Computer Assisted Intervention (MICCAI).

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], and Melanie Ganz[1,3]

[1] Neurobiology Research Unit, Rigshospitalet, Copenhagen (CPH), Denmark (DK)
[2] Faculty of Health and Medical Sciences, University of Copenhagen, CPH, DK
[3] Department of Computer Science, University of Copenhagen, CPH, DK
[4] Department of Biostatistics, University of Copenhagen, CPH, DK

Abstract. Brain imaging studies have the potential to predict treatment effects on neurotransmitters and receptors in the living human brain following a pharmacological intervention. However, data arising from neuroimaging 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. Hence, there is a rising concern about the validity and reproducibility of scientific studies in general [1] and especially in neuroimaging [2, 3].

Data sharing initiatives such as OpenNeuro (openneuro.org) are now enabling researchers to open up the subject selection and data acquisition factors of a 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). 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 [4, 5].

Conversely, the influence of the preprocessing on the outcome of the data analysis has besides a few initiatives in fMRI [5, 6] 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[5] that is used daily to extract features from subjects enrolled in neuroscience research studies. Furthermore, 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.
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 [3]. 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. Furthermore, there might not even exist a single best workflow for all studies of a given type. Indeed, there is evidence that different workflows might be optimal for different studies or even for different individuals [6]. 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 you have preprocessed your data - would be able to detect the effect. Hence, it is also of interest to analyze not only the variance in the preprocessing [5, 6], 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 and 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, $\boldsymbol{x}$, with class labels y, to generalize to the true underlying joint distribution $p(\boldsymbol{x}, y)$. Our approach adopts a range of preprocessing choices as a generative model for $\boldsymbol{x}$, and evaluates the predictive performance for the conditional distribution $p(y|\boldsymbol{x})$ using permutations [7] and the max statistic [11]. By permuting 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 [8].

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 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 **(C)** estimation of the statistical significance of the prediction accuracy (Figure 1).

---

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

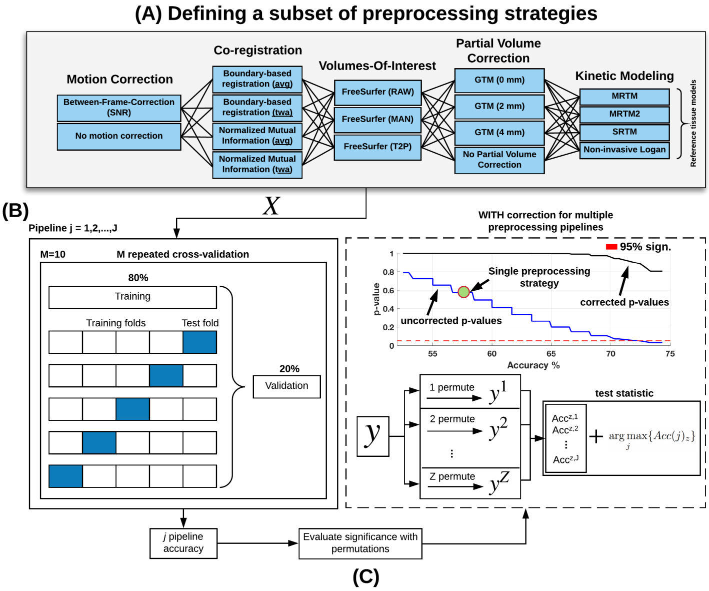

Fig. 1: **(A)** Definition of a subset of preprocessing strategies $j = 1, ..., J$: This includes preprocessing steps such as motion correction, co-registration, delineation of volumes of interest, partial volume correction, and kinetic modeling. **(B)** Model selection and cross-validation: For each pipeline j, select a classification model (e.g. Linear Discriminant), and a nested cross-validation scheme with M repetitions, 80% training data, and 20% validation data. **(C)** Evaluate 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 preprocessing pipelines and for Z permutations, generate a null-distribution of maximal accuracies across preprocessing pipelines. 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 randomly permuted null-distribution at a significance level α.

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, where $\boldsymbol{x_n} \in \mathbb{R}^p$ are observations with p features and $y_n \in \{-1, 1\}$ are the corresponding class labels, the data is commonly preprocessed using a set of steps such as motion correction, co-registration and partial volume correction (Figure 1A). The entire sequence of preprocessing steps is often referred to as a pipeline, designed to remove artifacts and noise from the data. Designing the most optimal sequence of 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 model and tune the model parameters to the data, and (2) assess the chosen predictive model by estimating the future prediction ability of the model. For both (1) and (2), one common approach is to use cross-validation and evaluate the model in an independent test set (Figure 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) using e.g. 5-fold cross-validation. The validation data has to be independent of the training data and completely held out of the training procedure. Additionally, the procedure has to be repeated so that each observation is assigned to the validation data exactly once. 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 [9].

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 (Figure 1C). The significance of each model and pipeline is estimated by randomly permuting the class labels Z times (i.e. sampling a permutation π^z from a uniform distribution over the set, $\mathbf{\Pi}_N$, of all permutations of indices $1, ..., N$) and re-running the above M times repeated 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 maximal accuracies across 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 given preprocessing pipeline j, we compute the permutation distribution of the maximal 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 maximal accuracy across pipelines J

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

- Construct an empirical cumulative distribution of max 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 actual labels for each pipeline j, $t_{0,j} = Acc(\mathbf{x}_{1,j}, y_{1,j}, ..., \mathbf{x}_{N,j}, y_N)$, and its corresponding p-value p^j_0 under the empirical distribution $\hat{P}_{max}$.

The null hypothesis assumes that the two classes have identical distributions,

$$\forall \boldsymbol{x} : p(\boldsymbol{x}|y = 1) = p(\boldsymbol{x}|y = -1).$$

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 maximal accuracy. We can reject the null hypothesis for any preprocessing pipeline with an accuracy exceeding this threshold.

2.5 Use of the max statistic in neuroimaging

Correction of p-values using the maximal statistic has been used before in statistical studies of neuroimaging data [10, 11]. Furthermore, several studies have examined the effects of multiple preprocessing options in combination with prediction [5, 6]. 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 a single experiment: a longitudinal PET study with a baseline and a re-scan in 31 healthy participants, following a pharmacological intervention [8]. The data, $\mathbf{x}$, consists of 60 observations (29 paired observations, 1 baseline, and 1 intervention) each with levels of serotonin transporter binding ($\mathrm{BP_{ND}}$, [12]) in 34 cortical brain regions covering the entire neocortex. For quantification of $\mathrm{BP_{ND}}$, 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 $\mathrm{BP_{ND}}$ (MRTM, SRTM, Non-invasive Logan and MRTM2). This results in $2 \times 3 \times 4^3 = 384$ combinations of preprocessing. Details are described in [13]. In the experiment, we used a Linear Discriminant to train a classifier to predict the classes (baseline and intervention), and jackknifing (i.e., sampling without replacement) for cross-validation. The number of cross-validation iterations was 10, and the number of permutation iterations was 1,000. To obtain true independence between the data and the labels in the

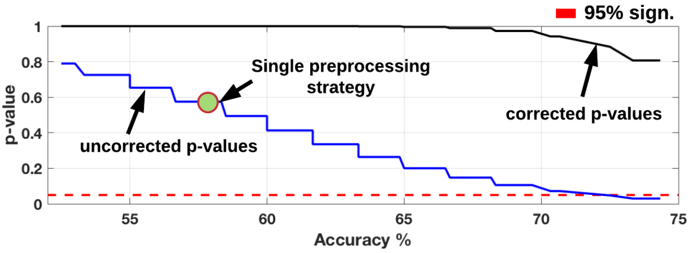

Fig. 2: Accuracy (%) as a function of p-value for 384 preprocessing strategies. The blue line indicates the p-values according to their permuted null distribution (uncorrected) and the black line indicates the p-values according to the maximal permuted null distribution (corrected). The red dotted line is 95% significance level.

cross-validation, observations for each subject (i.e. baseline and intervention) were always together.

We start by studying the behaviour of accuracies and p-values, when varying the preprocessing strategy, reported in Figure 2. Every point on the blue line and the black line is a preprocessing strategy with an accuracy and a p-value, respectively. 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 Figure 2 shows the p-values relative to the maximal 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 maximal permuted (blue). Most of preprocessing strategies fall within the permuted null distribution, but a subset of 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). To reject the null hypothesis under the empirical distribution of the maximal classification accuracies across pipelines, one would need an expected classification accuracy of 85% to obtain statistical significance at $\alpha = 0.05$ (Figure 2).

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 with the majority of preprocessing choices leading to a non-significant prediction (uncorrected). When correcting the significant pipelines using knowledge about all the applied pipelines, no significant predictions survive (corrected using the max statistics).

While the statistical analysis of each individual preprocessing pipeline is done in an optimal fashion due to the use of nested cross-validation, some of the preprocessing

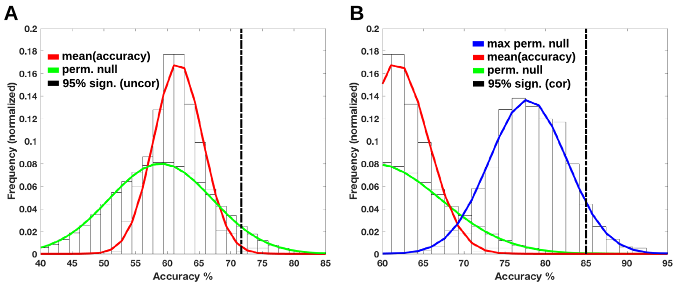

Fig. 3: **(A)** Average classification accuracies across preprocessing pipelines obtained using nested cross-validation with 10 repeats (red). The permuted null distribution of classification accuracies (1000 permutations) across preprocessing pipelines is visualized by the green distribution. The vertical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

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 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 proposed framework is 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 predictive models with varying model complexities. However, 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 [3], but should also provide the variance of their results across many different 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.

References

1. Andrew Gelman and Eric Loken. The statistical crisis in science. American Scientist, 2014, 102(6):460-65.
2. Button, K. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5):365-376.

3. Poldrack, R. A., et al. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. Nature Reviews Neuroscience, 18(2), 115-126.
4. Eklund, A., Nichols, T. E., Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 113(28):7900-7905.
5. Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. Frontiers in Neuroscience, 6(OCT), 1-13.
6. Churchill, N. W., et al. (2015). An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. PLoS ONE, 10(7), 1-25.
7. Golland, P & Fischl, B. Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies. Information Processing in Medical Imaging (IPMI), 2003, 18:330-41.
8. Frokjaer, V. G., 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.
9. Varoquaux G, et al. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage. Volume 145, Part B, 15 January 2017, Pages 166-179.
10. Holmes, A. P., et al. (1996). Nonparametric analysis of statistic images from functional mapping experiments. JCBFM, 16(1), 7-22.
11. Nichols, T. E., Holmes, A. P. (2001). Nonparametric Permutation Tests for PET functional Neuroimaging Experiments: A Primer with examples. HBM, 15(1), 1-25.
12. Innis, R. B., et al. (2007). Consensus nomenclature for in vivo imaging of reversibly binding radioligands. JCBFM, Sep;27(9):1533-9.
13. Nørgaard, M., et al (2018). The Impact of Preprocessing Pipeline Choice in Univariate and Multivariate Analyses of PET Data. IEEE Xplore (PRNI 2018), 1-4.

APPENDIX F

Declarations of Co-Authorship

GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

DECLARATION OF CO-AUTHORSHIP

Information on PhD student:	
Name of PhD student	Martin Nørgaard
E-mail	martin.noergaard@nru.dk
Date of birth	19/04/1990
Work place	Neurobiology Research Unit, Copenhagen University Hospital
Principal supervisor	Prof. Gitte Moos Knudsen

Title of PhD thesis:

Optimizing Preprocessing Pipelines in PET/MR Neuroimaging

This declaration concerns the following article:

Cerebral Serotonin Transporter Measurements with [11C]DASB: A Review on Acquisition and Preprocessing Across 21 PET Centres

The PhD student's contribution to the article: ***(please use the scale (A,B,C) below as benchmark*)***	**(A,B,C)**
1. Formulation/identification of the scientific problem that from theoretical questions need to be clarified. This includes a condensation of the problem to specific scientific questions that is judged to be answerable by experiments	C
2. Planning of the experiments and methodology design, including selection of methods and method development	C
3. Involvement in the experimental work	C
4. Presentation, interpretation and discussion in a journal article format of obtained data	C

****Benchmark scale of the PhD student's contribution to the article***		
A. refers to:	***Has contributed to the co-operation***	***0-33 %***
B. refers to:	***Has contributed considerably to the co-operation***	***34-66 %***
C. refers to:	***Has predominantly executed the work independently***	***67-100 %***

Signature of the co-authors:

Date:	**Name:**	**Title:**	**Signature:**
08/01/19	Melanie Ganz	Asst. Prof.	Melanie Ganz-Benjaminsen
8/1-19	Claus Svarer	Senior scientist	
14/1-19	Gitte Moos Knudsen	Professor	

Signature of the PhD student and the principal supervisor:	
Date: 11/1-2019 **PhD student:**	**Date:** 14/1-19 **Principal supervisor:**

GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

DECLARATION OF CO-AUTHORSHIP

Information on PhD student:	
Name of PhD student	Martin Nørgaard
E-mail	martin.noergaard@nru.dk
Date of birth	19/04/1990
Work place	Neurobiology Research Unit, Copenhagen University Hospital
Principal supervisor	Prof. Gitte Moos Knudsen

Title of PhD thesis:

Optimizing Preprocessing Pipelines in PET/MR Neuroimaging

This declaration concerns the following article:

Optimization of Preprocessing Strategies in Positron Emission Tomography (PET) Neuroimaging: A [11C]DASB Study

The PhD student's contribution to the article: *(please use the scale (A,B,C) below as benchmark*)*	**(A,B,C)**
1. Formulation/identification of the scientific problem that from theoretical questions need to be clarified. This includes a condensation of the problem to specific scientific questions that is judged to be answerable by experiments	C
2. Planning of the experiments and methodology design, including selection of methods and method development	B
3. Involvement in the experimental work	A
4. Presentation, interpretation and discussion in a journal article format of obtained data	C

****Benchmark scale of the PhD student's contribution to the article***		
A. refers to:	***Has contributed to the co-operation***	***0-33 %***
B. refers to:	***Has contributed considerably to the co-operation***	***34-66 %***
C. refers to:	***Has predominantly executed the work independently***	***67-100 %***

Signature of the co-authors:

Date:	**Name:**	**Title:**	**Signature:**
08/01/19	Melanie Ganz	Asst. Prof.	Melanie Ganz-Benjaminsen
8/1-19	Claus Svarer	Senior scientist	
24/1.19	Vibe G. Frokjaer	Senior scientist	
1/17/19	Douglas N. Greve	Asst. Prof.	DNGreve

21/1 – 19	Stephen C. Strother	Professor	S. C. Strother.
23/1 -19	Gitte Moos Knudsen	Professor	
[redacted]			

Signature of the PhD student and the principal supervisor:	
Date: 11/1 - 2019	**Date:** 23/1-19
PhD student:	**Principal supervisor:**

GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

DECLARATION OF CO-AUTHORSHIP

Information on PhD student:	
Name of PhD student	Martin Nørgaard
E-mail	martin.noergaard@nru.dk
Date of birth	19/04/1990
Work place	Neurobiology Research Unit, Copenhagen University Hospital
Principal supervisor	Prof. Gitte Moos Knudsen

Title of PhD thesis:

Optimizing Preprocessing Pipelines in PET/MR Neuroimaging

This declaration concerns the following article:

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

The PhD student's contribution to the article: *(please use the scale (A,B,C) below as benchmark*)*	(A,B,C)
1. Formulation/identification of the scientific problem that from theoretical questions need to be clarified. This includes a condensation of the problem to specific scientific questions that is judged to be answerable by experiments	C
2. Planning of the experiments and methodology design, including selection of methods and method development	C
3. Involvement in the experimental work	C
4. Presentation, interpretation and discussion in a journal article format of obtained data	C

**Benchmark scale of the PhD student's contribution to the article*		
A. refers to:	*Has contributed to the co-operation*	*0-33 %*
B. refers to:	*Has contributed considerably to the co-operation*	*34-66 %*
C. refers to:	*Has predominantly executed the work independently*	*67-100 %*

Signature of the co-authors:

Date:	Name:	Title:	Signature:
08/01/19	Melanie Ganz	Asst. Prof.	Melanie Ganz-Benjaminsen
8/1-19	Claus Svarer	Senior scientist	
23/1-19	Gitte Moos Knudsen	Professor	
21/01-19	Stephen C. Strother	Professor	S. C. Strother

1/17/19	Douglas N. Greve	Asst. Prof.	DNGreve
24.1.19	Vibe G. Frøkjær	Senior scientist	

Signature of the PhD student and the principal supervisor:

Date: 11/1-2019	**Date:** 14/1-19
PhD student:	**Principal supervisor:**

GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

DECLARATION OF CO-AUTHORSHIP

Information on PhD student:	
Name of PhD student	Martin Nørgaard
E-mail	martin.noergaard@nru.dk
Date of birth	19/04/1990
Work place	Neurobiology Research Unit, Copenhagen University Hospital
Principal supervisor	Prof. Gitte Moos Knudsen

Title of PhD thesis:
Optimizing Preprocessing Pipelines in PET/MR Neuroimaging

This declaration concerns the following article:
The Impact of Different Preprocessing Strategies in PET Neuroimaging: A [11C]DASB-PET Study

The PhD student's contribution to the article: *(please use the scale (A,B,C) below as benchmark*)*	(A,B,C)
1. Formulation/identification of the scientific problem that from theoretical questions need to be clarified. This includes a condensation of the problem to specific scientific questions that is judged to be answerable by experiments	C
2. Planning of the experiments and methodology design, including selection of methods and method development	C
3. Involvement in the experimental work	C
4. Presentation, interpretation and discussion in a journal article format of obtained data	C

**Benchmark scale of the PhD student's contribution to the article*		
A. refers to:	*Has contributed to the co-operation*	*0-33 %*
B. refers to:	*Has contributed considerably to the co-operation*	*34-66 %*
C. refers to:	*Has predominantly executed the work independently*	*67-100 %*

Signature of the co-authors:			
Date:	Name:	Title:	Signature:
08/01/19	Melanie Ganz	Asst. Prof.	Melanie Ganz Benjaminsen
8/1-19	Claus Svarer	Senior scientist	
23/1-19	Gitte Moos Knudsen	Professor	
21/01-19	Stephen C. Strother	Professor	S. C. Strother

1/17/19	Douglas N. Greve	Asst. Prof.	DNGreve

Signature of the PhD student and the principal supervisor:	
Date: 11/1 - 2019 **PhD student:**	**Date:** 14/1 - 19 **Principal supervisor:**

GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

DECLARATION OF CO-AUTHORSHIP

Information on PhD student:	
Name of PhD student	Martin Nørgaard
E-mail	martin.noergaard@nru.dk
Date of birth	19/04/1990
Work place	Neurobiology Research Unit, Copenhagen University Hospital
Principal supervisor	Prof. Gitte Moos Knudsen

Title of PhD thesis:

Optimizing Preprocessing Pipelines in PET/MR Neuroimaging

This declaration concerns the following article:

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

The PhD student's contribution to the article: ***(please use the scale (A,B,C) below as benchmark*)***	**(A,B,C)**
1. Formulation/identification of the scientific problem that from theoretical questions need to be clarified. This includes a condensation of the problem to specific scientific questions that is judged to be answerable by experiments	C
2. Planning of the experiments and methodology design, including selection of methods and method development	C
3. Involvement in the experimental work	C
4. Presentation, interpretation and discussion in a journal article format of obtained data	C

****Benchmark scale of the PhD student's contribution to the article***		
A. refers to:	***Has contributed to the co-operation***	***0-33 %***
B. refers to:	***Has contributed considerably to the co-operation***	***34-66 %***
C. refers to:	***Has predominantly executed the work independently***	***67-100 %***

Signature of the co-authors:

Date:	**Name:**	**Title:**	**Signature:**
25/1/19	Melanie Ganz	Asst. Prof.	Melanie Ganz
25/1-19	Claus Svarer	Senior scientist	Claus Svarer
24/1-19	Vibe G. Frøkjær	Senior Scienist	

22/01	Brice Ozenne.	Post-Doc	

Signature of the PhD student and the principal supervisor:	
Date: 22/1 - 2019 **PhD student:**	**Date:** 23/1 -19 **Principal supervisor:**