Original Article

JCBFM

Journal of Cerebral Blood Flow & Metabolism
2020, Vol. 40(9) 1902–1911

DOI: 10.1177/0271678X19880450
journals.sagepub.com/home/jcbfm
SAGE

Different preprocessing strategies lead to different conclusions: A [^{11}C]DASB-PET reproducibility study

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

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 commonly apply different arrays of sequential data analytic methods ("preprocessing pipeline"), but it is often unknown how the choice of preprocessing 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 affects the outcome of the study. We tested 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. The two preprocessing steps that were most critical for the outcome were motion correction and kinetic modeling of the dynamic PET data. We found that 36% of the applied preprocessing strategies replicated the originally reported finding ($p < 0.05$). For preprocessing strategies with motion correction, the replication percentage was 72%, whereas it was 0% for strategies without motion correction. In conclusion, the choice of preprocessing strategy can have a major impact on a study outcome.

Keywords

Positron emission tomography, preprocessing, head motion, partial volume correction, kinetic modeling

Received 14 June 2019; Revised 26 August 2019; Accepted 7 September 2019

Introduction

Science is entering a reproducibility crisis.[1] Historically, this has meant being unable to reproduce scientific results in an independent sample, even when using the same experimental design and methodological choices.[2]

In practice, the outcome of two similar studies is never 100% overlapping because of differences in methodology, e.g. available equipment, settings, and sample data.[3]

Apart from differences in methodology, 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.[4]

In positron emission tomography (PET) neuroscience, only a few studies have investigated the impact of methodological choices on the outcome of a study.

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

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

Samper-González et al.[5] 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.[6] 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.[7,8]

Mukherjee et al.[9] 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.[10,11] More recently, Nørgaard et al.[12] 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 carried out 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.[13–16]

Recently, we showed that inconsistent reports of 5-HTT levels in healthy individuals might be explained by variations in acquisition and preprocessing strategy.[17]

However, while it may be inevitable that different methods are applied in different PET centres, the key question that remain unanswered 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.[18] which is a double blind, randomized, placebo-controlled intervention study of 60 healthy women. We applied 384 different preprocessing strategies to test their sensitivity to reproduce the main outcome from Frokjaer et al.,[18] namely a positive association between the emergence of depressive symptoms and change in cerebral 5-HTT binding following a pharmacological sex-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.[18] Because preprocessing strategies in the [^{11}C]DASB-PET literature have been assumed to produce near similar results,[19–21] 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

Participants

A total of 60 female participants (mean age 24.3 ± 4.9 years) were included in a double-blind, randomized, placebo-controlled study,[18] 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 Supplementary Table S1 and in Frokjaer et al.[18] The study was registered and approved by the ethics committee for the capital region of Copenhagen (protocol-ID: H-2-2010-108) and registered as a clinical trial: www.clinicaltrials.gov under the trial ID NCT02661789. All subjects provided written informed consent prior to participation, in accordance with The Declaration of Helsinki II.

Positron emission tomography

All participants were scanned in a Siemens ECAT HRRT scanner with the selective 5-HTT radioligand [^{11}C]DASB.[22] The protocol consisted of a 90-min dynamic acquisition (3D list-mode) post injection of 587 ± 30 (mean $\pm$ SD) MBq bolus into an elbow vein. The PET data were reconstructed into 36 frames (6×10, 3×20, 6×40, 5×60, 5×120, 8×300, 3×600 s) using a 3D-OSEM-PSF algorithm with TXTV attenuation correction.[23,24]

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.

Magnetic resonance imaging

An isotropic T1-weigthed MP-RAGE was acquired for all participants (matrix size $= 256 \times 256 \times 192$; voxel size $=$ 1 mm; TR/TE/TI $= 1550/3.04/800$ ms; flip angle $= 9^\circ$) using either a Siemens Magnetom Trio 3T or a Siemens 3T Verio MR scanner. Furthermore, an isotropic T2-weighted sequence (matrix size $256 \times 256 \times 176$; voxel size $= 1$ mm; TR/TE $= 3200/409$ ms; flip angle $= 120^\circ$) was acquired for all participants. All acquired MRI's were corrected for gradient nonlinearities,[25] and examined to ensure the absence of structural abnormalities.

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), PVC and kinetic modeling) with each step consisting of two to four choices.

All preprocessing strategies have previously been applied and evaulated.[17] The steps are listed below in the order in which they were applied, producing a total of 384 different preprocessing strategies (Figure 1). The outcome measure for each preprocessing strategy is an estimate of the brain regional non-displaceable binding potential (BP_{ND}).[26]

Further details on all preprocessing steps can be found in Nørgaard et al.[17]

Motion correction (two choices). The PET data were 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.[18] 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 participants had acceptable motion below 3 mm.

Co-registration (four choices). 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)[27] or boundary-based registration (BBR)[28] were subsequently applied to either the twa or the avg image.

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

Delineation of VOI (three choices). All MRI's were processed (recon-all) using FreeSurfer (http://surfer.nmr.mgh.harvard.edu, version 5.3).[29] 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.[18] The neocortex region was generated by taking all cortical TACs in the Desikan-Killiany 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

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

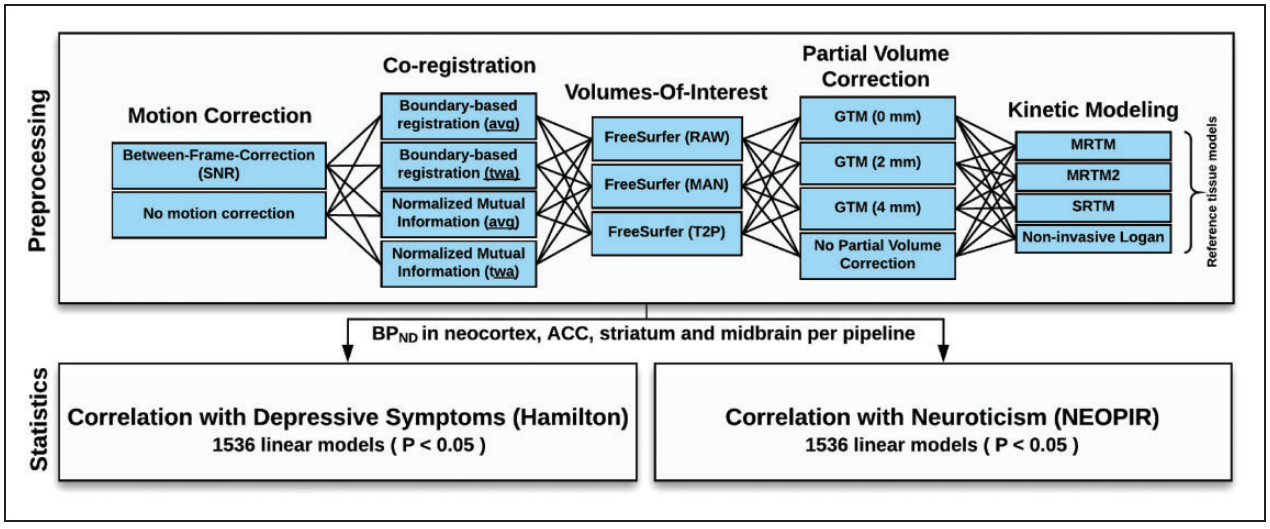

Figure 1. Schematic overview of the 384 preprocessing strategies applied for the [^{11}C]DASB quantification. The output from the preprocessing is the non-displaceable binding potential (BP_{ND}) in the regions neocortex, anterior cingulate cortex (ACC), striatum and midbrain, which are subsequently entered into the statistical analysis, including the correlation with depressive symptoms (Hamilton) and the neuroticism (NEOPIR). This sums to a total of $384 \times 4 \times 2 = 3072$ statistical tests (significance level, $p < 0.05$). avg: average; twa: time-weighted average; SNR: signal-to-noise ratio; GTM: geometric transfer matrix.

The striatum was generated by averaging the regions putamen and caudate.[30] The remaining regions, ACC and midbrain, were automatically generated by FS.

Partial volume correction (four choices). The PET data were corrected either without (noPVC) or with PVC. The VOI-based PVC technique, Geometric Transfer Matrix (GTM) by Rousset et al.,[31] was applied using PETsurfer (surfer.nmr.mgh.harvard.edu/fswiki/PetSurfer)[6] 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,[32] 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.

Kinetic modeling (four choices). Four kinetic models were applied, all based on reference tissue modeling (RTM) and implemented in MATLAB 2016b (https://www.mathworks.com). 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.[33] The non-invasive Logan reference tissue model was applied as described in Logan et al.[34] The simplified reference tissue model (SRTM) was applied as described in Lammertsma and Hume.[35] For MRTM2 and non-invasive Logan, the thalamus, putamen and caudate were averaged to represent a single less noisy high-binding region for estimation of k_2' using MRTM. All kinetic models applied in this work were implemented in MATLAB as specified in their original paper, and fitted using the weighting scheme

$$w = \sqrt{\frac{scan\ duration^2}{total\ counts\ in\ frame}}.$$

The implementation in MATLAB was validated with PMOD v. 3.0 (10 subjects < 0.1% difference in BP_{ND}), but was carried out in MATLAB for parallel execution purposes to substantially reduce processing time.

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 × 2 dependent variables × 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 (uncorrected). The rationale for excluding the correction of *p*-values due to the use of multiple preprocessing strategies (as should otherwise always be carried out in post-hoc analyses) is because we wanted to make our analysis as comparable as possible to the original study. Secondly, it was not our primary goal to question the results from Frokjaer et al.,[18] but instead to simulate the situation, had another preprocessing strategy been used in the original study. The main question was to investigate the sensitivity of the results to different analysis pipelines, driven by the following assumptions and/or starting point: (1) researchers within a PET center often use a local preprocessing strategy, that consists of the steps: MC, co-registration, delineation of VOIs, PVC, kinetic modeling, and (2) it is assumed that each PET center only applies a single preprocessing strategy, hence there is no need to correct for multiple preprocessing strategies.

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.

Depressive symptoms and change in [^{11}C]DASB binding from baseline across preprocessing strategies

The sensitivity of motion correction on the *p*-values from the association between Hamilton change from baseline and change in neocortical BP_{ND} from baseline is shown in Figure 2. *P*-values obtained with strategies using MC varied between 0.014 and 0.091, with 72% of strategies falling below the 0.05 significance boundary. *P*-values obtained without MC varied between 0.091 and 0.44, with 0% of strategies falling below the 0.05 significance boundary. Across all 384 preprocessing strategies, 36% of *p*-values were below 0.05. Effect sizes (i.e. Pearson's correlation) varied from 0.15 to 0.45 (Figure S1).

The *p*-values also marginally depended on which kinetic model was applied after MC, with non-invasive Logan and MRTM2 resulting in lower *p*-values, compared to the corresponding preprocessing strategies using MRTM and SRTM (Figure S3(c)).

Table 1. Baseline levels of binding across preprocessing strategies.

	Placebo ($n = 30$)	GnRHa ($n = 29$)	GnRHa versus placebo p-value	% preprocessing strategies with $p < 0.05$
Neocortex	$0.98 \pm .46$	$0.95 \pm .46$	$0.25 \pm .12$	0
ACC	$1.39 \pm .46$	$1.32 \pm .46$	$0.17 \pm .60$	0.5
Striatum	$2.69 \pm .34$	$2.51 \pm .35$	$0.21 \pm .12$	11.5
Midbrain	$2.27 \pm .34$	$2.24 \pm .36$	$0.73 \pm .18$	0

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

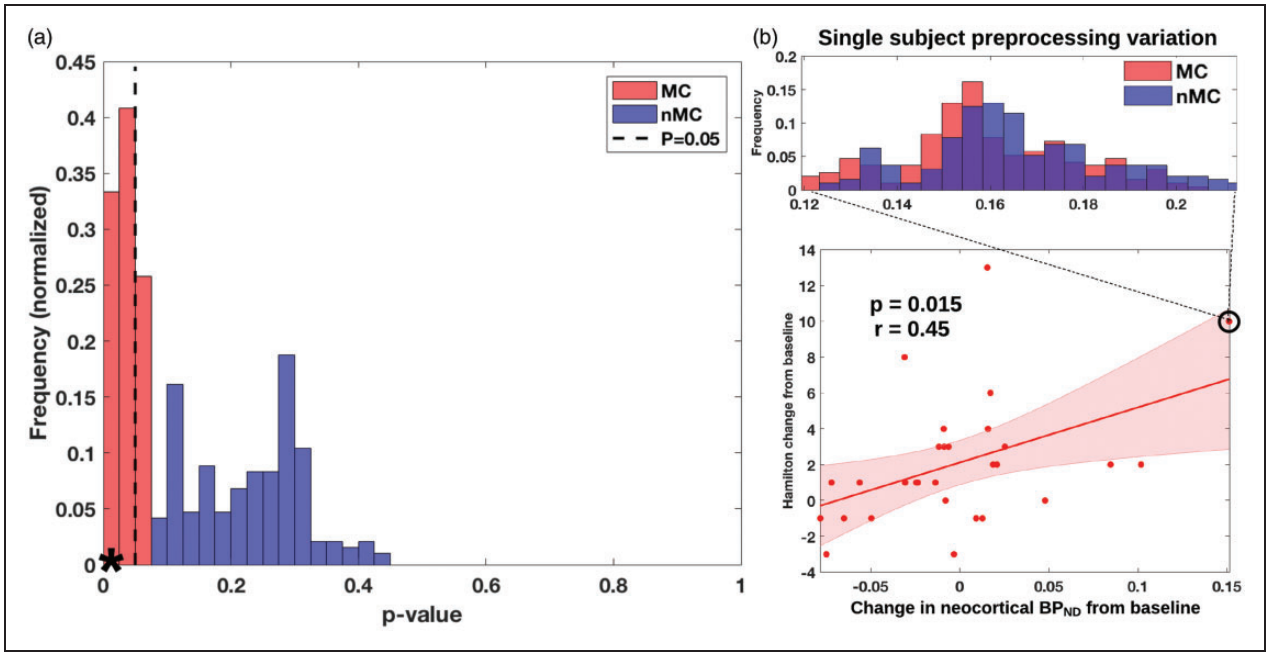

Figure 2. (a) Histogram of p-values obtained across 384 preprocessing strategies examining the association between change in neocortical BP_{ND} and change in Hamilton score from baseline in the GnRHa group. MC: motion correction; nMC: no motion correction; SRTM: simplified reference tissue model. (b) 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.[17] (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 (with MC = 72%, without MC = 0%). 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.

PVC with GTM generally increased the p-values with increasing PSF (0 mm, 2 mm, 4 mm), with this effect being most evident when MC was applied (Figure S3(b)).

Neither the techniques for co-registration nor delineation of VOIs caused any consistent differences on the resulting p-values (Figure S3(d) and S3(e)).

We also identified how the change in neocortical BP_{ND} from baseline varied across preprocessing strategies for a single subject (Figure 2(b), upper). Notably, the size of this variability (range: 0.12–0.22) was similar in size to the between-subject variability ranging from −0.07 to 0.1 (Figure 2(b), lower).

The remaining histograms, raw p-values and estimates for k_2' as a function of preprocessing strategy, can be found in the Supplementary Material.

Neuroticism and 5-HTT binding across preprocessing strategies

Figure 3 shows the evaluation of preprocessing strategies as a function of p-value for the association between neuroticism and change in ACC BP_{ND} from baseline, ranging from 0.014 to 0.59. Across all preprocessing strategies, 93% failed to identify a significant

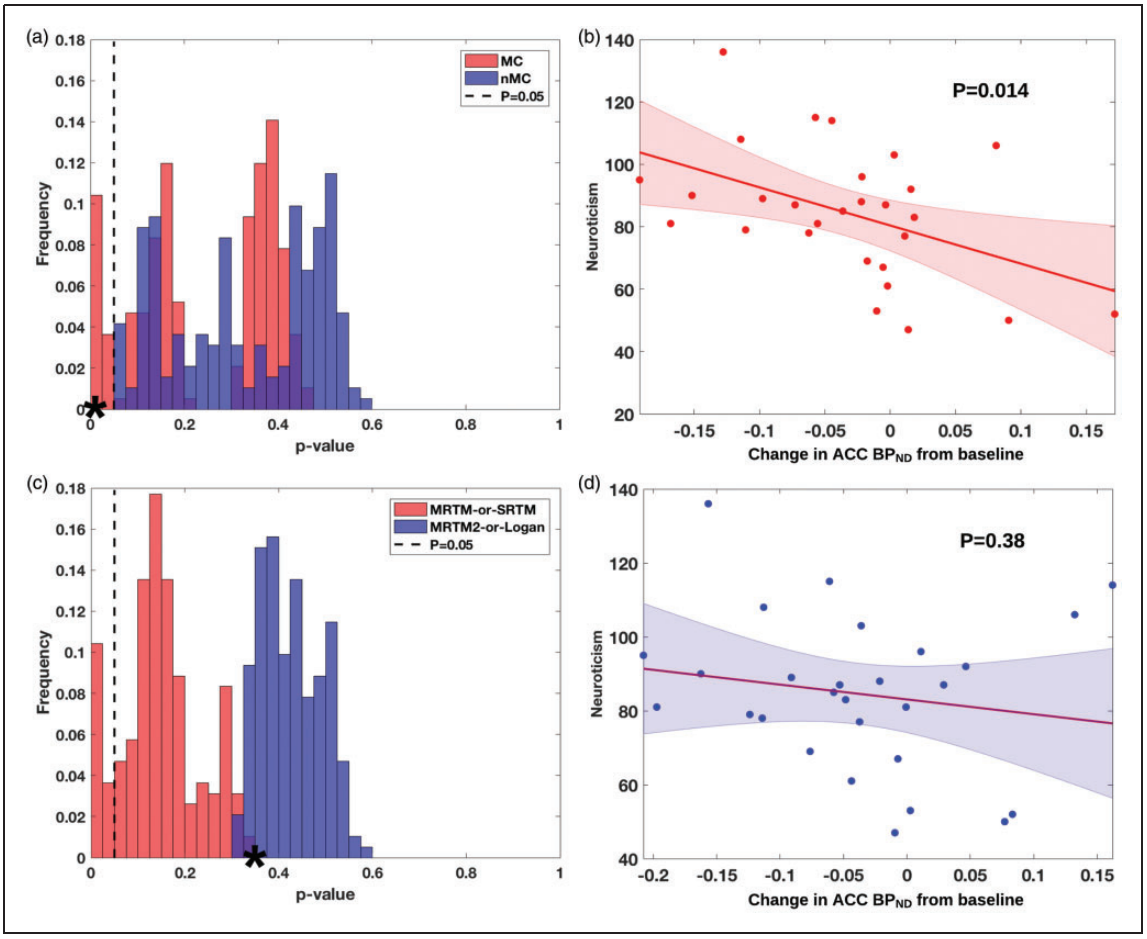

Figure 3. (a) Histogram of obtained *p*-values for the association between the change in ACC BP_{ND} from baseline and neuroticism, in the GnRHa group and across 384 preprocessing strategies. MC: motion correction; nMC: no motion correction. (b) Association between the increase in ACC BP_{ND} from baseline 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)). MC: motion correction; SRTM: simplified reference tissue model; MRTM: multilinear reference tissue model; ACC: anterior cingulate cortex.

association, whereas 7% of strategies containing MC and MRTM/SRTM produced a *p*-value below the 0.05 significance boundary. The application of MRTM/SRTM versus Logan/MRTM2 showed a clear segregation in *p*-values, with Logan/MRTM2 ranging between 0.3 and 0.59, and MRTM/SRTM ranging between 0.01 and 0.34.

The supplementary material contains *p*-values from all 3072 linear regression models in freely available MATLAB files (*.mat). All the reported data are available through the CIMBI database.[36]

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.

While small variations in preprocessing strategy between studies have generally been considered to be insensitive to the outcome, we identified several preprocessing steps having an impact on the outcome. One of the most notable observations of our analysis was that MC had a major impact on the replication of the original results, with the absence of MC leading to a 0% replication despite varying the remaining preprocessing

steps. Various approaches for co-registration and delineation of VOIs resulted in only minor effects on the main outcome; this is not unexpected because the preprocessing steps are carried out in subject space. By contrast, had the data been registered to and VOIs delineated in standard space both the bias and the variance would have been affected,[17] potentially leading to more noisy estimates.

PVC led to only marginally higher *p*-values which might be explained by a PVC related slight increase in between-subject variability.

While kinetic modeling had only minor effects on the sensitivity to detect the association between depressive symptoms and neocortical BP_{ND} (Figure S3(c)), we found that choosing SRTM-or-MRTM versus MRTM2-or-Logan had a clear impact on the *p*-values for the association between neuroticism and ACC BP_{ND} (Figure 3). The kinetic models mainly differ from each other in the model parameter estimation, whether they are linear (MRTM2-or-Logan)[33] or non-linear (SRTM).[35] In addition, biological assumptions and noise control also differ (e.g. MRTM vs. MRTM2) as well as the number of parameters that need to be estimated, i.e. three parameters for SRTM-or-MRTM and two parameters for MRTM2-or-Logan. This means that there will be a bias-variance trade-off to consider, as a reduction in model parameters to fit the data will reduce the variance of the model, at the expense of a bias.[33]

Finally, the estimated k_2' will be different between preprocessing strategies, but will also have different impact depending on the brain region of interest. In a post hoc analysis of the variability of k_2' across preprocessing strategies, we identified a marginally lower k_2' with MC compared to without MC (Figure S4(a)), and an increase in k_2' with an increase in PSF using GTM (Figure S4(b)). Ichise et al.[33] reported that a positive bias in k_2' will lead to a negative bias in the BP_{ND}. Furthermore, if the noise in the signal is increased, e.g. without MC and/or following the application of PVC, it will positively bias and increase the variability of the k_2' estimate, further negatively biasing the estimate of the BP_{ND}.[33] The bias in k_2' is expected to be small because we used an average of striatum (putamen and caudate) and thalamus to generate a single, less noisy high-binding region for estimation of k_2'. In addition, despite higher subject and/or preprocessing dependent noise levels will increase the variability of k_2', we find it unlikely that differences in k_2' between preprocessing strategies explain the observed differences in kinetic modeling outcomes (Figure 3). Instead, in the presence of high noise levels, MRTM2 will be less subject to higher variability and more bias in BP_{ND} as compared to MRTM-or-SRTM, likely explaining the differences in sensitivity we are observing for kinetic modeling. Since the true noise and estimates for k_2' and BP_{ND} are unknown, it is difficult to disentangle the contribution from each subject on the group correlation structure in Figure 2. This is because the error in the estimated value for each subject will introduce a bias in all BP_{ND} estimates for that subject.[33]

Another notable observation was that the single-subject variability resulting from preprocessing strategy was nearly as large as the between-subject variability (Figure 2(b), 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.[18] 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 3). While neuroticism has consistently been implicated in stress regulation, depression and brain 5-HTT,[30,37] 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.[38–40] In general, acute stress enhances serotonin output, and in turn, serotonin signaling influences the secretion of corticosteroids.[19,41] 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.[12,19] 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 S2 in the supplementary) for some but not all regions and preprocessing strategies (*p*-values

< 0.05, uncorrected). 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 neuroticism dependent variation between baseline and 5-HTT binding and is in line with other observations.[42] 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 a 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.[43] 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.[44]

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 if all pipelines are considered, 72% confidence if pipelines with MC are considered, and 0% if pipelines without MC are considered). 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.[45,46] 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.,[12] 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 are tested using one single framework, for the effect of MC using the AIR 5.3.0 package[47] 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.

Conclusions

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.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: MN was supported by the National Institutes of Health (Grant 5R21EB018964-02), the Lundbeck Foundation (Grant R90-A7722), the Innovation Fund Denmark (4108-00004B), and the Independent Research Fund Denmark (DFF-1331-00109 & DFF-4183-00627). MG was supported by the Lundbeck Foundation (R181-2014-3586). DNG was supported by National Institutes of Health (Grants R01 EB023281, 5R01NS083534, 5R01EB019956, 1R01NS105820).

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. Finally, we thank the anonymous reviewers for their careful reading and complimentary comments that helped improve this paper.

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.

Authors' contributions

VGF acquired the data. MN, MG, CS, VGF, DNG, SCS, GMK analyzed the data. MN drafted the manuscript, and MG, CS, VGF, DNG, SCS and GMK revised and contributed to the final version.

Supplemental material

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

ORCID iD

Martin Nørgaard https://orcid.org/0000-0003-2131-5688

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016; 533: 452–454.
2. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2016; 349: aac4716.
3. Goodman SN, Fanelli D and Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med* 2016; 8: 341ps312.
4. Strother SC, Anderson J, Hansen LK, et al. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage* 2002; 15: 747–71.
5. Samper-González J, Burgos N, Bottani S, et al. 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; 183: 504–521.
6. Greve DN, Salat DH, Bowen SL, et al. Different partial volume correction methods lead to different conclusions: an 18F-FDG-PET study of aging. *Neuroimage* 2016; 132: 334–343.
7. Berkouk K, Quarantelli M, Prinster A, et al. Mapping the relative contribution of gray matter activity vs. volume in brain PET: a new approach. *J Neuroimaging* 2006; 16: 224–235.
8. Meltzer CC, Zubieta J, Brandt J, et al. Regional hypometabolism in Alzheimer's disease as measured by PET following correction for effects of partial volume averaging. *Neurology* 1996; 47: 454–461.
9. Mukherjee JM, Lindsay C, Mukherjee A, et al. Improved frame-based estimation of head motion in PET brain imaging. *Med Phys* 2016; 43: 2443–2454.
10. 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 Imaging* 2013; 32: 200–209.
11. Anton-Rodriguez JM, Sibomana M, Walker MD, et al. Investigation of motion induced errors in scatter correction for the HRRT brain scanner. In: *IEEE Nuclear Science Symposuim & Medical Imaging Conference*, Knoxville, TN, 30 October-6 November 2010, pp. 2935–2940. IEEE.
12. Nørgaard M, Ganz M, Svarer C, et al. Cerebral serotonin transporter measurements with [11C]DASB: a review on acquisition and preprocessing across 21 PET Centres. *J Cereb Blood Flow Metab* 2019; 39: 210–222.
13. Frick A, Åhs F, Engman J, et al. Serotonin synthesis and reuptake in social anxiety disorder. *JAMA Psychiatry* 2015; June: E1–E9.
14. Zientek F, Winter K, Moller 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: 2460–2466.
15. Kim E, Howes OD, Park JW, et al. Altered serotonin transporter binding potential in patients with obsessive-compulsive disorder under escitalopram treatment: [11C]DASB PET study. *Psychol Med* 2016; 46: 357–366.
16. Hinderberger P, Rullmann M, Drabe M, et al. 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* 2016; 110: 530–536.
17. Nørgaard M, Ganz M, Svarer C, et al. Optimization of preprocessing strategies in positron emission tomography (PET) neuroimaging: a [11C]DASB study. *Neuroimage* 2019; 199: 466–479.
18. 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 Psychiatry* 2015; 78: 534–543.
19. Kim JS, Ichise M, Sangare J, et al. PET imaging of serotonin transporters with [11C]DASB: test-retest reproducibility using a multilinear reference tissue parametric imaging method. *J Nucl Med* 2006; 47: 208–214.
20. 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.
21. 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.
22. Houle S, Ginovart N, Hussey D, et al. Imaging the serotonin transporter with positron emission tomography: initial human studies with [11C]DAPP and [11C]DASB. *Eur J Nucl Med* 2000; 27: 1719–1722.
23. 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: 1000–1008.
24. Keller SH, Svarer C and Sibomana M. Attenuation correction for the HRRT PET-scanner using transmission scatter correction and total variation regularization. *IEEE Trans Med Imaging* 2013; 32: 1611–1621.
25. Jovicich J, Czanner S, Greve DN, et al. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage* 2006; 30: 436–443.
26. Innis RB, Cunningham VJ, Delforge J, et al. Consensus nomenclature for in vivo imaging of reversibly binding radioligands. *J Cereb Blood Flow Metab* 2007; 27: 1533–1539.
27. Studholme C, Hill DLG and Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recogn* 1999; 32: 71–86.
28. Greve D and Fischl B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 2008; 48: 63–72.
29. Fischl B. FreeSurfer. *Neuroimage* 2012; 62: 774–781.
30. 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.
31. Rousset OG, Ma Y and Evans AC. Correction for partial volume effects in PET: principle and validation. *J Nucl Med* 1998; 39: 904–911.
32. Olesen OV, Sibomana M, Keller SH, et al. Spatial resolution of the HRRT PET scanner using 3D-OSEM PSF reconstruction. In: *IEEE Nucl Scie Symp Conf Record (NSS/MIC)*, Orlando, FL, 24 October-1 November 2009, pp. 3789–3790. IEEE.
33. Ichise M, Liow J-S, Lu J-Q, et al. Linearized reference tissue parametric imaging methods: application to [11C]DASB positron emission tomography studies of the serotonin transporter in human brain. *J Cereb Blood Flow Metab* 2003; 23: 1096–1112.
34. 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.
35. Lammertsma AA and Hume SP. Simplified reference tissue model for PET receptor studies. *Neuroimage* 1996; 4: 153–158.
36. Knudsen GM, et al. The center for integrated molecular brain imaging (Cimbi) database. *Neuroimage* 2016; 124: 1213–1219.
37. Hirvonen J, Tuominen L, Någren K, et al. Neuroticism and serotonin 5-HT1Areceptors in healthy subjects. *Psychiatry Res* 2015; 234: 1–6.
38. 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.
39. 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.
40. Jakobsen GR, Fisher PM, Dyssegaard, et al. Brain serotonin 4 receptor binding is associated with the cortisol awakening response. *Psychoneuroendocrinology* 2015; 67: 124–132.
41. Lanfumey L, Mongeau R, Cohen-Salmon C, et al. Corticosteroid-serotonin interactions in the neurobiological mechanisms of stress-related disorders. *Neurosci Biobehav Rev* 2008; 32: 1174–1184.
42. Stenbæk DS, Budtz-Jørgensen E, Pinborg, et al. Neuroticism modulates mood responses to pharmacological sex hormone manipulation in healthy women. *Psychoneuroendocrinology* 2019; 99: 251–256.
43. Gabrieli JDE, Ghosh SS and Whitfield-Gabrieli S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 2015; 85: 11–26.
44. Yarkoni T and Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci* 2017; 12: 1100–1122.
45. Greve DN and Fischl B. False positive rates in surface-based anatomical analysis. *Neuroimage* 2017; 171: 6–14.
46. Benjamin DJ, Berger JO and Johnson VE. Commentary: redefine statistical significance. *Nat Hum Behav* 2018; 2: 6–10.
47. Woods RP, Cherry SR and Mazziotta JC. Rapid automated algorithm for aligning and reslicing PET images. *J Comput Assist Tomograph* 1992; 16: 620–33.