Magia: Robust automated modeling and image processing toolbox for PET neuroinformatics

Tomi Karjalainen[1*], Severi Santavirta[1], Tatu Kantonen[1], Jouni Tuisku[1], Lauri Tuominen[1,2], Jussi Hirvonen[1,3], Jarmo Hietala[1], Juha Rinne[1], and Lauri Nummenmaa[1,4]

[1]Turku PET Centre, University of Turku, Finland

[2]Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[3]Department of Radiology, University of Turku, Finland

[4]Department of Psychology, University of Turku, Finland

***Corresponding author**

Tomi Karjalainen

Turku PET Centre

c/o Turku University Hospital

P.O. Box 52

20520 Turku, Finland

Email: tomi.karjalainen@utu.fi

Keywords: PET, neuroinformatics, modelling, neuroimaging

Short title: Automated reference tissue modelling

Competing Interests' Statement: None

Acknowledgements: This work was supported by the Academy of Finland grants #265915 and #294897 to LN and Sigrid Juselius Foundation grant to LN, Päivikki and Sakari Sohlberg Foundation to ToK, Finnish Cultural Foundation Varsinais-Suomi Regional Fund to ToK

Abstract

Introduction: Processing PET data typically requires substantial manual labour including data retrieval, input function processing, drawing reference regions, performing realignment, and kinetic modelling. It is thus not suited for large-scale standardized reanalysis as processing one single image may take many hours to process. To resolve this problem, we introduce the Magia pipeline for brain-PET and MRI data that enables automatic processing of PET data with minimal user intervention. Here we investigated the accuracy of Magia in the automatic brain-PET data processing with four tracers binding to different binding sites: [11C]raclopride, [11C]carfentanil, [11C]madam, and [11C]PiB.

Materials and methods: For each tracer, we analyzed 30 historical control subjects' data with manual and automated methods. In the manual method, five persons delineated the reference regions (cerebellum and occipital cortex) for each image according to written instructions. The automatic ROI extraction was based on FreeSurfer parcellations. Data were modelled using simplified reference tissue yielding voxelwise BPnd values, except for [11C]PiB where ROI-to-cerebellum ratio was used. We compared the anatomical overlap between automatically and manually defined reference regions. We also computed the similarity of reference time-activity curves (TAC), and similarity of outcome measures at voxel and ROI level for both techniques.

Results: Anatomically, MAGIA-generated reference regions differed from manual reference regions. The mean overlap percentage per tracer ranged from 20.3% (SEM 1.70) for [11C]raclopride to 7.3% (SEM 0.28) for [11C]madam. Mean overlap percentages for [11C]carfentanil and [11C]PiB were 8.1% (SEM 0.92) and 8.0% (SEM 0.25), respectively. Reference region TACs and results from outcome measures are demonstrated in Fig 1. and Fig 2.

Conclusion: Our results confirm that although the automated reference regions are anatomically different from manually drawn regions, the correlation coefficient of time-activity curves is over 0.99 and the resulting outcome measures are comparable. Based on these results and taking account the lack of subjectivity in automatic reference region delineation, the high level of standardization and strong scalability, we conclude that MAGIA-pipeline can efficiently process the brain-PET data for at least these four tracers.

Introduction

Statistical power of neuroimaging studies has been widely questioned in the recent years, leading to the conclusions that significantly larger samples are required for avoiding false positive and negative findings (Button et al., 2013; Cremers et al., 2017; Yarkoni, 2009). Additionally, the role of researcher degrees of freedom, i.e. the subjective choices made during the process from data collection to its analysis, has been identified as an important reason for poor replicability of many findings. (Simmons et al., 2011) Consequently, the focus in neuroimaging has shifted towards standardized, large-scale neuroinformatics based approaches (Poldrack & Yarkoni, 2016; Yarkoni et al., 2011). Today, several standardized and highly automatized preprocessing pipelines are publicly available for processing functional magnetic resonance images (Esteban et al., 2019). Such standardized methods are not, however, currently widely used for analysis of positron emission tomography (PET) data.

The primary bottleneck for automatization of PET analysis is the requirement of input function. Depending on tracer, the input function can be obtained either from blood samples or directly from the PET images if a reference region is available for the tracer. The blood samples require substantial manual processing before the input function can be obtained from them. While population-based atlases (Eickhoff et al., 2005; Fischl et al., 2002; Tzourio-Mazoyer et al., 2002) provide an automatic way for defining reference regions (Schain et al., 2014; Tuszynski et al., 2016; Yasuno et al., 2002), they are suboptimal because the process requires spatial normalization of the images. Ideally, the reference region should be defined separately for each individual before spatial normalization. Thus, the golden standard method for defining the reference region is still its manual delineation. The delineation process is time-consuming and relies on several subjective choices. To minimize between-study variance resulting from operator-dependent choices (White et al., 1999), a single individual should delineate the reference regions for all studies within a project. Thus, manual delineation is not suited for large-scale projects where hundreds of scans are processed, or neuroinformatics approaches where even significantly larger number of scans have to be processed.

To resolve these problems, we have introduce the Magia analysis pipeline for brain-PET data that enables automatic modelling of PET data with minimal user intervention (https://github.com/tkkarjal/magia). The major advantages of this approach involve:

1) Flexible, parallelizable environment suitable for large-scale standardized analysis
2) Fully automated processing of PET data from raw image files to uptake estimates.
3) Visual and quantitative quality control of the processing steps.

4) Centralized management and storage of study metadata, image processing methods and outputs for subsequent reanalysis and quality control.

In this study we tested the reliability of the automatic reference region generation, input function extraction, modelling, and spatial preprocessing of PET data with four tracers with different binding sites: [11C]raclopride, [11C]carfentanil, [11C]MADAM, and [11C]PiB by comparing the Magia-derived input functions and uptakes against those obtained using conventional manual techniques. We also assessed inter-rater agreement in the reference region definition and uptake estimates, and regional and voxel-level outcome measures.

Materials and methods

Magia

Magia (https://github.com/tkkarjal/magia) is a fully automatic analysis pipeline running on MATLAB. It combines methods from SPM (www.fil.ion.ucl.ac.uk/spm/) and FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) as well as in-house software developed for modeling PET data. Magia has been developed alongside a centralized database containing metadata about each study, facilitating data storage and neuroinformatics-type large-scale PET analyses. However, Magia can also be installed and used without such database. Magia requires MATLAB, SPM, and FreeSurfer and runs on Linux/Mac. The Optimization Toolbox for MATLAB is required for fitting the ROI level models. Magia has been developed using MATLAB R2016b. Given a detailed description of a brain PET study, Magia automatically chooses one of eight alternative analysis branches to process the study. The way a study is processed depends on if the study in question is dynamic or static, if an MRI is available, and if plasma input is available. Magia currently recognizes 10 tracers, each of which have their own default modeling method with default modeling parameters. Magia currently supports the simplified reference tissue model (SRTM), Patlak with both plasma input and reference tissue input, SUV-ratio for both dynamic and static studies, and FUR analysis for late scans with plasma input. If the tracer is not recognized, Magia proceeds by calculating standardized uptake values for the study.

A box-diagram describing the main steps in Magia processing is shown in **Figure 1**. Magia starts by preprocessing the PET images. This includes frame-alignment and coregistration with the MRI. The MRI is processed with FreeSurfer to generate anatomical parcellations for defining regions of

interest (Schain et al., 2014), including a reference region if one is available for a tracer. Magia performs a two-step correction to the reference tissue mask (see below) before obtaining the input-function for modeling; the corrections make the reference region generation robust for many scanners and individuals. The MRI is also segmented into volumometric grey and white matter probability maps for spatial normalization (Ashburner & Friston, 2000). After modelling, the obtained parametric images are normalized and smoothed. In addition to the parametric images, Magia also calculates ROI level parametric estimates for each study. Finally, the results are stored in a centralized archive in a standardized format, facilitating future population-level analyses with visual and quantitative quality control metrics.

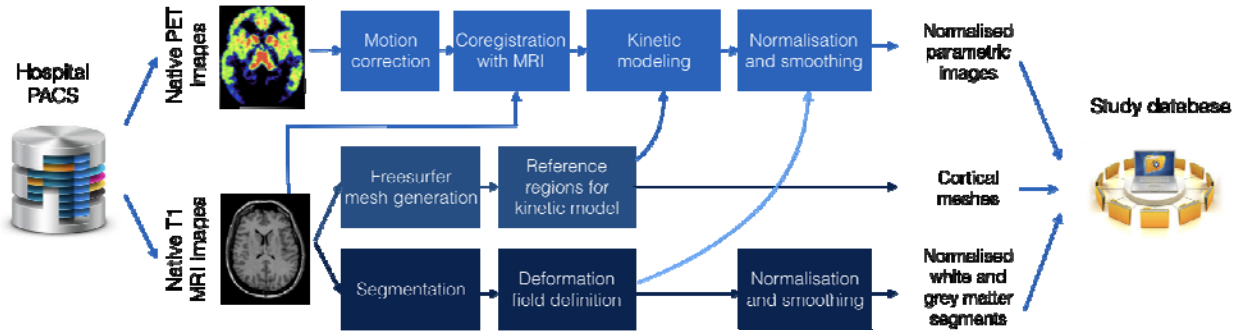

Figure 1. *The MAGIA pipeline combining FreeSurfer cortical mesh generation and parcellation, T1 MRI image segmentation and normalization, automatic reference region and region of interest generation, and kinetic modeling.*

Figure 1b.

Above-mentioned steps are only used when applicable. For example, for static images the frame alignment is skipped, and if there is no related MRI available, then a tracer-specific template must be provided to normalize the images. Magia also supports tracers that do not have a reference region. For such studies, the preprocessed plasma input must be available.

Validation data

To assess reliability of Magia we used historical control data using four radioligands with different targets and spatial distribution of inding sites: Dopamine D2R receptor ligand [11C]raclopride, mu-opioid receptor ligand [11C]carfentanil, serotonin transporter ligand [11C]MADAM, and beta-amyloid ligand [11C]PIB. For each radioligand we selected 30 studies (Table 1). We generated

reference regions for all the tracers using traditional manual methods and the new automatic method and compared the results.

	[^{11}C]carfentanil	[^{11}C]raclopride	[^{11}C]MADAM	[^{11}C]PiB
N (female)	30 (12)	30 (23)	30 (17)	30 (18)
Age (mean, range)	32 (20 - 51)	39 (20 - 60)	42 (25 - 57)	71 (66 - 80)
Scanners	HRRT PET/CT PET/MR	GE Advance PET/CT HRRT	HRRT	HRRT
Data range (years)	2007 - 2016	1998 - 2014	2008 - 2015	2014 - 2016

Table 1. Summary of the studies. Scanners: HRRT (HRRT, Siemens Medical Solutions); PET/CT (Discovery 690 PET/CT, GE Healthcare); PET/MR (Ingenuity TF PET/MR, Philips Healthcare); GE Advance (GE Advance, GE Healthcare).

Manual reference region delineation

Five researchers with good knowledge of human neuroanatomy delineated reference regions for every study according to written and visual instructions (**Figure 2a**). Cerebellum was used as a reference region for [11C]raclopride (Gunn et al., 1997), [11C]MADAM (Lundberg et al., 2005) and [11C]PiB (Lopresti et al., 2005). For [11C]carfentanil, occipital cortex was used (Endres et al., 2003). The regions were drawn using CARIMAS (http://turkupetcentre.fi/carimas/). The reference regions were defined on three consecutive transaxial T1-weighted MR images, which is the current standard method at Turku PET Centre. Cerebellar reference was drawn in cerebellar gray matter within a gray zone in the peripheral part of cerebellum, distal to the bright signal of white matter. The first cranial slice was placed below occipital cortex to avoid spill-in of radioactivity. Typically, this is a slice where the temporal lobe is clearly separated from the cerebellum by the petrosal part of the temporal bone. The most caudal slice was typically located in the most caudal part of the cerebellum. Laterally, venous sinuses were avoided to avoid spill-in during early phases of the scans. Posteriorly, there was about a 5 mm distance from cerebellar surface to avoid spill-out effects. Anteriorly, the border of the reference region was drawn approximately 2 mm distal to the border or

cerebellar white and gray matter, except in the most caudal, where central white matter may no longer be visible.

The occipital reference region was defined on three consecutive transaxial slices, of which the most caudal slice was the second-most caudal slice before cerebellum. The reference region was drawn J-shaped with medial and posterior parts. The reference region was drawn to roughly follow the shape of the cortical surface, but not individual gyri. The reference region was drawn approximately 1 cm wide with about 2 mm margin to the cortical surface to avoid spill-out effects. The anterior border of the reference region was placed approximately halfway between the posterior cortical surface and the splenium of corpus callosum. The posterolateral border of the reference region approximated the medial-most part of the posterior horn of the lateral ventricle.

Automatic reference region generation

Figure 2b shows an overview of the autoamted ROI definition process. First, T1-weighted MR images were fed into FreeSurfer to provide study-specific reference regions. Second, an anatomical correction was applied to the FreeSurfer-generated reference region mask to remove voxels that, based on their anatomical location alone, were the most likely to suffer from spillover effects or that might contain also specific binding. For cerebellum, the most important sources of spillover effects are occipital cortex and venous sinuses. Thus, the most outermost cerebellar voxels were excluded in the anatomical reference region correction. For occipital cortex, voxels that were lateral to the lateral ventricles were excluded. This is because the most lateral parts of the FreeSurfer-generated occipital cortex extend to areas with specific binding for [11C]carfentanil, and the lateral ventricles provide an easy and reliable reference point for thresholding. Finally, the radioactivity concentration distribution within the anatomically corrected reference region was estimated, and the tails of the distribution aere excluded. The lower and upper boundaries for the signal intensities are defined by calculating the full width at half maximum (FWHM) of the mean PET signal intensity distribution and excluding voxels that are on the tail-ends of the corresponding radioactivity concentrations. This step ensures that the reference region will not contain voxels with atypically high or low signal, and thus reflect the typical values for unspecific binding. Thus, the automatic reference region generation process combines information from anatomical brain scans and the PET images.

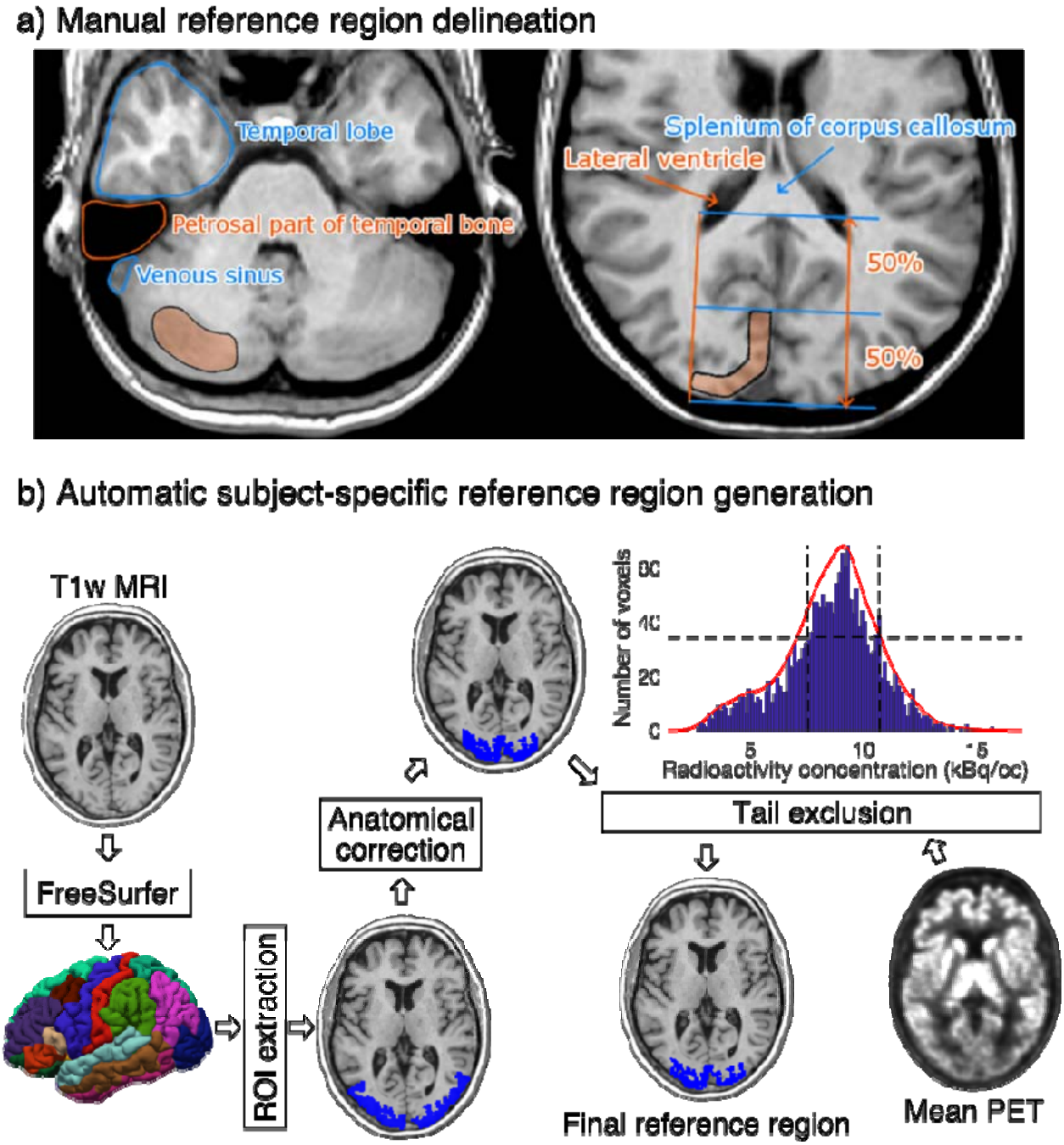

Figure 2. *a) Visual instructions of the most cranial slice of manually delineated cerebellar (left) and occipital (right) reference regions. The reference regions were delineated on three consecutive transaxial T1-weighted MR images. Cerebellar reference region is shown on the left and occipital reference region on the right. b) The diagram shows how a T1-weighted magnetic resonance image of an individual's brain is processed to produce the final reference region. The shown example is from the [11C]carfentanil data set. The rectangles represent processing steps between inputs and outputs. The FreeSurfer step assigns an anatomical label to each voxel of the subject's T1 weighted MR image. The ROI extraction step extracts a prespecified region of interest from FreeSurfer's output. The anatomical correction*

removes voxels that are most likely to suffer from spillover effects; in [11C]carfentanil data this means excluding voxels lateral to the lateral ventricles. In the tail exclusion step, a PET signal intensity distribution within the anatomically corrected reference region is defined, and the voxels whose intensities are on the tail-ends of the distribution are excluded from the reference region.

Quantifying operator-dependent variability

We first quantified operator-dependent variability on the reference regions, input functions and outcome measures. Within-study overlap between the manual reference regions was used to quantify *anatomical* similarity of the reference regions. The overlap was first calculated separately for all different manual reference region pairs and then the mean overlap was assessed for each study. Pearson correlation coefficient and AUC were used to compare reference region time-activity curves. Pearson correlations for every manual reference region pair was calculated, and their median was used to index within-study similarity. We also investigated whether outcome measures (BPnd/SUVR) differed between manually delineated reference regions. To assess similarity of AUCs and outcome measures, we conducted all pairwise comparisons between individually drawn reference regions.

Volumetric similarity of the manual and automatic reference regions

We compared the volumes of reference regions to assess whether the two techniques generate reference regions of systematically different sizes. For each study, we calculated the mean volume from all manually delineated reference regions and compared it to volume of the Magia-derived reference region. We also quantified the anatomical overlap between the manually and the automatically derived reference regions. The overlap was defined as ratio between the number of common voxels and the number of manual voxels. For each study, the overlap was first calculated separately for every manually delineated reference region and then the mean overlap was assessed.

Similarity of the reference region radioactivity concentrations

A functionally homogenous region should have approximately Gaussian distribution of radioactivity measured with PET (Teymurazyan et al., 2013). Functional homogeneousness was assessed using radioactivity distributions within the reference regions. The automatically and manually derived reference region masks were used to extract radioactivity concentration distribution within the reference regions. The study-specific manual distributions were averaged over the manual drawers to

provide a single manual distribution for each study. The radioactivity concentrations were converted into SUV, after which the distributions were averaged over studies to provide tracer-specific distributions. Mean, standard deviations, mode, and skewness of the distributions were used to quantify the differences in the distributions.

Similarity of the reference region time-activity curves

We compared the similarity of the automatically and manually delineated reference region time-activity curves (TACs). For each study, the manual reference region TAC was defined as the average across the manual TACs to minimize the subjective bias in adhering to the instructions for manual reference region delineation. Activities were expressed as standardized uptake values (SUV, g/ml) which were obtained by normalizing tissue radioactivity concentration (kBq/ml) by total injected dose (MBq) and body mass (kg), thus making the different images more comparable to each other. To assess the similarity of the shapes of reference region TACs, we calculated Pearson correlations between the manually and automatically delineated TACs for each tracer. Bias was assessed using area under curve (AUC).

Assessing similarity of the uptake estimates

We used nondisplaceable binding potential (BPnd) to quantify uptakes of [11C]carfentanil, [11C]raclopride and [11C]MADAM. It reflects the ratio between specific and nondisplaceable binding in the brain. The binding potentials were calculated using SRTM whose use has been validated for all tracers (Endres et al., 2003; Gunn et al., 1997; Lundberg et al., 2005). SUV-ratio was used to quantify [11C]PiB uptake (Lopresti et al., 2005). All the studies were first processed using Magia. The procedure was repeated with the only exception of replacing the automatically generated reference regions with a manually generated reference region. Thus, the only differences observed in the uptake estimates originate from differences in the reference regions. We calculated parametric images and also estimated the outcome measures in nine regions of interest (ROI) including both cortical and subcortical areas: amygdala, brainstem, caudate and thalamus as subcortical ROIs and medial orbitofrontal cortex (MOFC), superior temporal gyrus (STG) and postcentral gyrus (PCG) as cortical ROIs. We also used cerebellum as a ROI for [11C]carfentanil and lateral occipital cortex (LOC) as a ROI for [11C]raclopride, [11C]carfentanil, and [11C]MADAM. All ROIs were extracted from the FreeSurfer parcellations.

We wanted to assess how much variation in uptake estimates the subjective reference region delineation produces. For each tracer, we calculated the uptake estimates in a ROI with high specific binding. For every study, uptake was estimated using all the five manual reference regions and the Magia-derived reference region. Standard deviation of the tracer-specific uptake was used to assess the variation resulting from manual reference region delineation. While there were inter-individual differences in the means of the manual estimates, we assumed that the standard deviation is the same for all studies (homoscedasticity). Thus, the standard deviation estimates rely on 150 data points instead of 5. Finally, voxelwise outcome measure estimates were computed, and volumometric comparisons between the results from the two techniques were performed using SPM12.

Statistical analyses

Wilcoxon's matched pairs signed rank test was utilized for statistical comparison of reference region volumes, AUCs, and outcome measures. *P*-value of under 0.05 was considered statistically significant. Pearson correlation coefficient was used to assess differences in the shapes of the time-activity curves. All calculations and statistical analyses were executed using MATLAB R2016 (https://se.mathworks.com/products/matlab.html).

Results

Consistency of reference regions

We first compared the similarities between the automatically and manually delineated reference regions. For each tracer, automatic reference regions were, as expected, consistently larger than manually derived reference regions ($z > 4.35$, $p < 0.001$; **Figure 3a**). The median ratios between volumes of automatic and manual reference regions were approximately 2 (Q1–Q3: 1–2) for [11C]carfentanil, 3 (Q1–Q3: 2–4) for [11C]raclopride, 8 (Q1–Q3: 7–9) for [11C]MADAM and 8 (Q1–Q3: 7–9) for [11C]PiB. Four [11C]carfentanil studies had larger manual than automatic occipital reference regions (ratio from 0.67 to 0.99). Magia-generated cerebellar reference regions were always larger than mean manual cerebellar reference regions for all subjects and tracers.

Next, we determined whether automatically determined reference regions overlap with the manually drawn reference regions. Automatic occipital reference region for [11C]carfentanil overlapped only 14 % (Q1–Q3: 10.2–15.5) with manual occipital reference region. However, automatic cerebellar

reference regions overlapped manual reference regions by 55 %, 59 % and 61 % (Q1–Q3: 10–16, 51–60, 52–60, 57–68) for [11C]raclopride, [11C]MADAM and [11C]PiB, respectively. Overall *anatomically* automatic and manual reference regions were different, and the difference was not solely explained by the differences in their volumes. Additionally, the trimmed FreeSurfer-based reference region follows strictly the cortical grey matter surface spanning multiple transaxial slices in the image, whereas the manually drawn reference regions may contain significant amounts of white matter due to their intended expansion in x and y dimensions (see section Manual reference region delineation). Better overlap in cerebellar than occipital reference region was not surprising due to much larger ratio in volumes of cerebellar than occipital reference regions. We also investigated the overlap of manual reference regions between different operators drawing the ROIs for the study. As presented in **Figure 3b**, manual reference regions overlapped poorly between drawers. Tracer level median overlaps between drawers were 22 %, 41 %, 14 %, 18 %, for [11C]carfentanil, [11C]raclopride, [11C]MADAM and [11C]PiB, respectively. Poor overlap can be mostly explained by the fact that drawers often delineated the reference regions on different transaxial slices.

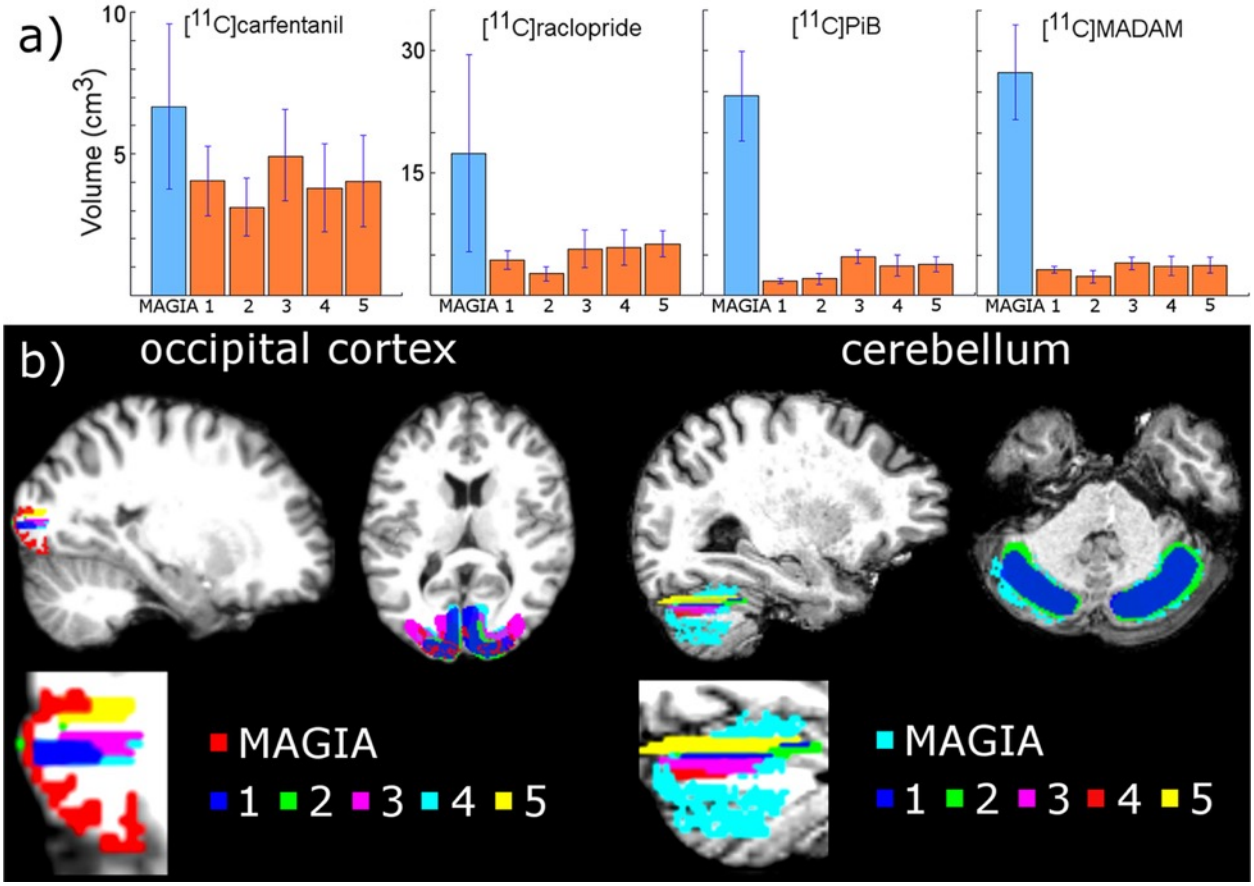

Figure 3. a) Mean volumes of MAGIA-generated reference regions compared to mean volumes of manually delineated reference regions. b) Visual example of MAGIA-generated and manual reference regions for one study.

Within-study variation in manually obtained reference tissue time-activity curves

The shapes of manual reference region time-activity curves were almost identical. The median Pearson correlation coefficient was over 0.99 for every tracer. Significant differences were observed between manual reference region AUCs. We conducted all pairwise comparisons of reference region AUCs and some, but not all, comparisons showed significant differences. The amount of significant pairwise comparisons are presented in parentheses for each tracer. For [11C]carfentanil (12/20) occipital cortex was the reference region and the median difference of significant pairwise comparisons of AUCs was 9 %. For [11C]raclopride (8/20), [11C]MADAM (10/20) and [11C]PiB (12/20) where cerebellum was the reference region median differences of significant comparisons of AUCs were 1 %, 2 %, 3 %, respectively.

Reference region SUV distributions

Mean reference region SUV distributions are shown in **Figure 4a** and time-activity curves of the reference regions in **Figure 4b.** The overlap between the manual and automatic distributions was approximately 90 % for all tracers. All distributions were unimodal and highly symmetric for all tracers. The means of the distributions were practically equal (maximum difference of 0.07 %). The standard deviations of the distributions differed by 14 %, 11 %, 12 % and 18% for [11C]carfentanil, [11C]MADAM, [11C]PIB and [11C]raclopride, respectively. The modes of the automatically and manually derived distributions were 1.5 and 1.55 for [11C]carfentanil, 1.95 and 2.05 for [11C]MADAM, 1.65 and 1.70 for [11C]PIB, and 1.35 and 1.35 for [11C]raclopride. Thus, the maximum difference was less than 5 %. The skewnesses of the Magia-derived and manually derived distributions were 1.2 and 0.9 for [11C]carfentanil (24 % difference), 1.3 and 1.2 for [11C]MADAM (11 % difference), 2.0 and 1.6 for [11C]PIB (26 % difference), and 2.4 and 2.0 for [11C]raclopride (21 % difference).

Reference region time-activity curves

The shapes of reference region time-activity curves were almost identical and the Pearson correlation coefficient (r) exceeded 0.99 for every tracer. AUCs were also highly similar. For [11C]carfentanil no statistically significant difference between automatic and manual AUC was observed. However, the difference between cerebellar reference region AUCs reached statistical significance. Automatic reference region AUCs for [11C]raclopride, [11C]MADAM and [11C]PiB were 2.7% ($p < 0.001$, Q1-Q3: 1.5%-4.7%), 2.4% ($p < 0.001$, Q1-Q3: 1.1%-3.3%) and 2.3% ($p < 0.001$, Q1-Q3: 0.0%-3.3%) smaller than manual reference region AUCs, respectively. Taken together, cerebellar reference region time-activity curves were slightly biased compared to manual reference region time-activity curves whereas no bias was observed for [11C]carfentanil.

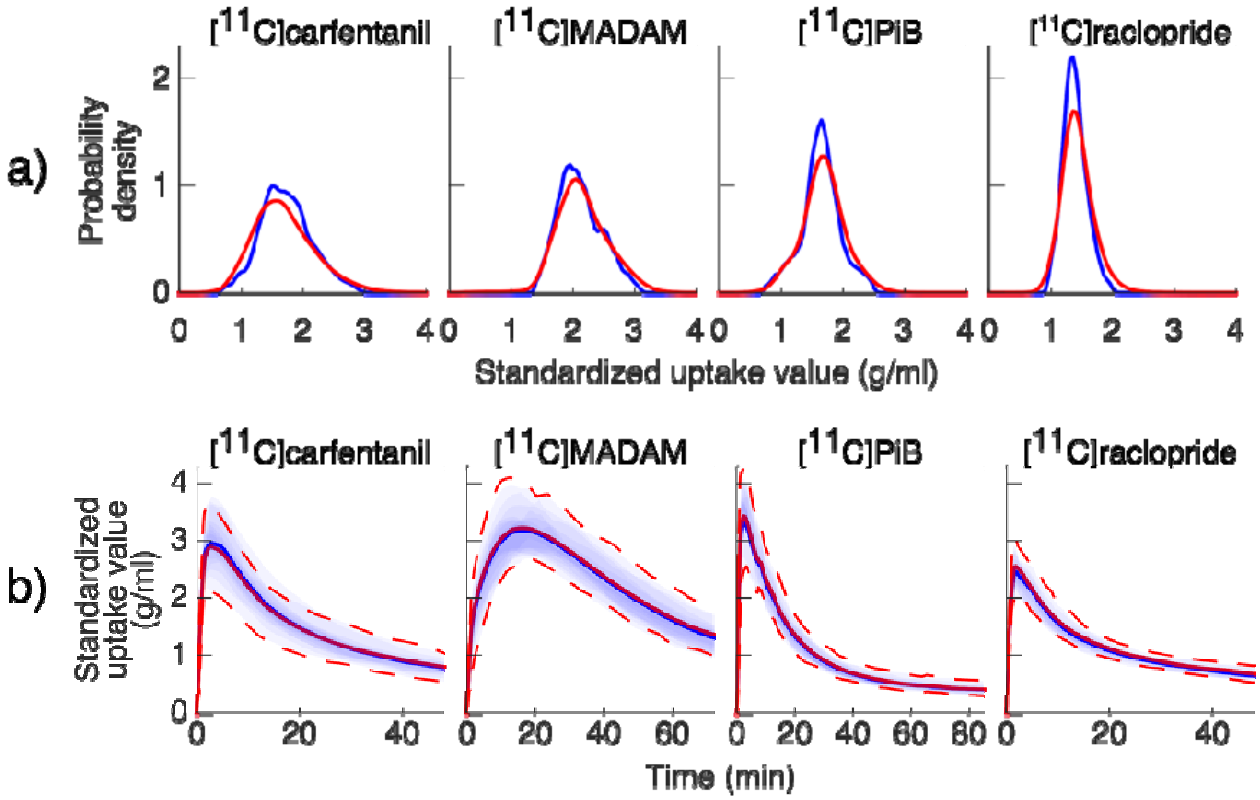

Figure 4. a) Probability density distributions of the standardized uptake values within the reference regions. b) Automatic and manual reference region time-activity curves and the respective 80 % percentile intervals. Blue = MAGIA, red = manual.

Similarity of the regional uptake estimates

Figure 5 shows how the Magia-derived outcome measures (A) differed from the average of the manual estimates that were regarded as the ground truth (B) and how these relate to bias (C). There was no systematic bias for [11C]carfentanil or [11C]PiB. For [11C]MADAM, Magia produced up to 3–5 % higher binding potential estimates in regions with high specific binding (**Figure 5B**). In cortical regions with low specific binding, the bias was over 10 %. For [11C]raclopride, Magia produced approximately 4–5 % higher binding potential estimates in striatum. In thalamus, the bias was 8–10 %. Elsewhere in the brain the bias varied considerably between 13–20 %. These differences were all statistically significant (FWE-corrected voxels, $p < 0.05$). For both [11C]MADAM and [11C]raclopride, the relative bias decreased significantly with increasing binding potential (**Figure 5C**).

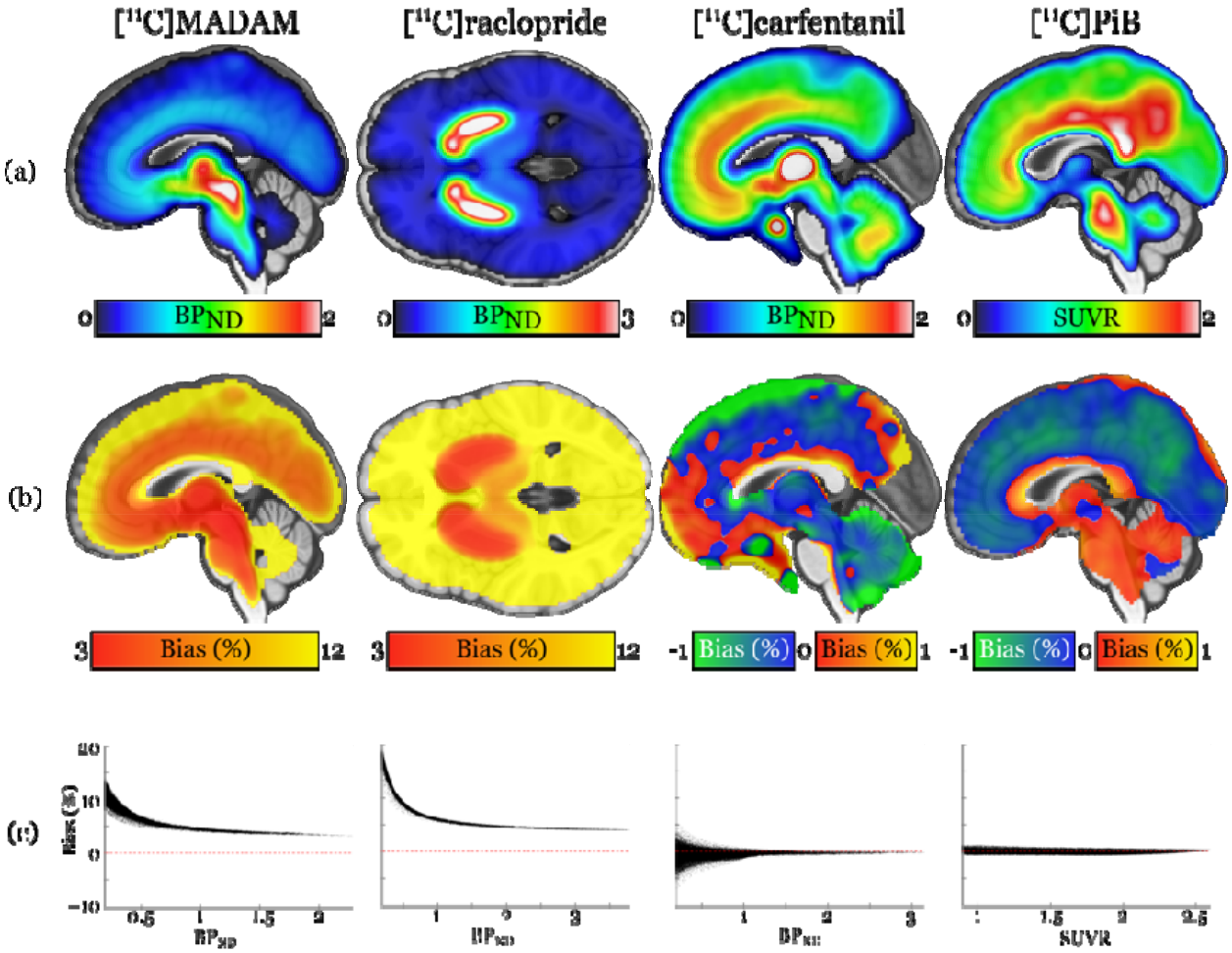

Figure 5. (a) Visualization of the outcome measure distributions for each tracer. (b) Maps visualizing the relative biases of the Magia-derived outcome measures compared to the averages

obtained by manual reference region delineation. The manual method is here presented as the ground truth, because the manual outcome for each scan is an average over five individual estimates, while the Magia result relies on a single estimate. (c) Associations between the outcome measure magnitude and relative bias.

Figure 6 shows the distributions of the mean outcome measures of each ROI for every tracer. In ROI-based analysis, there were no statistically significant differences of outcome measures in any ROI for [11C]carfentanil and [11C]PiB. However, significant differences were observed in every ROI for [11C]raclopride and [11C]MADAM. Magia produced up to 5% higher BPnd estimates for [11C]raclopride in caudate and putamen which are well-known high-binding areas. Notably, estimates were singificantly more variable in regions with low specific binding such as in cortex and brainstem for [11C] carfentanil (18-40% higher with MAGIA), possibly reflecting increased noise in the large regions with unspecific binding. Similarly, the bias in Magia produced BPnd estimates for [11C]MADAM were the lowest in high-binding areas (amygdala, thalamus, putamen) and the BPnd difference was up to 10% in these areas. The highest differences in BPnd estimates were observed in cortical low-binding areas (17-27%). Significant differences in outcome measures for [11C]raclopride and [11C]MADAM are shown in **Table 2.**

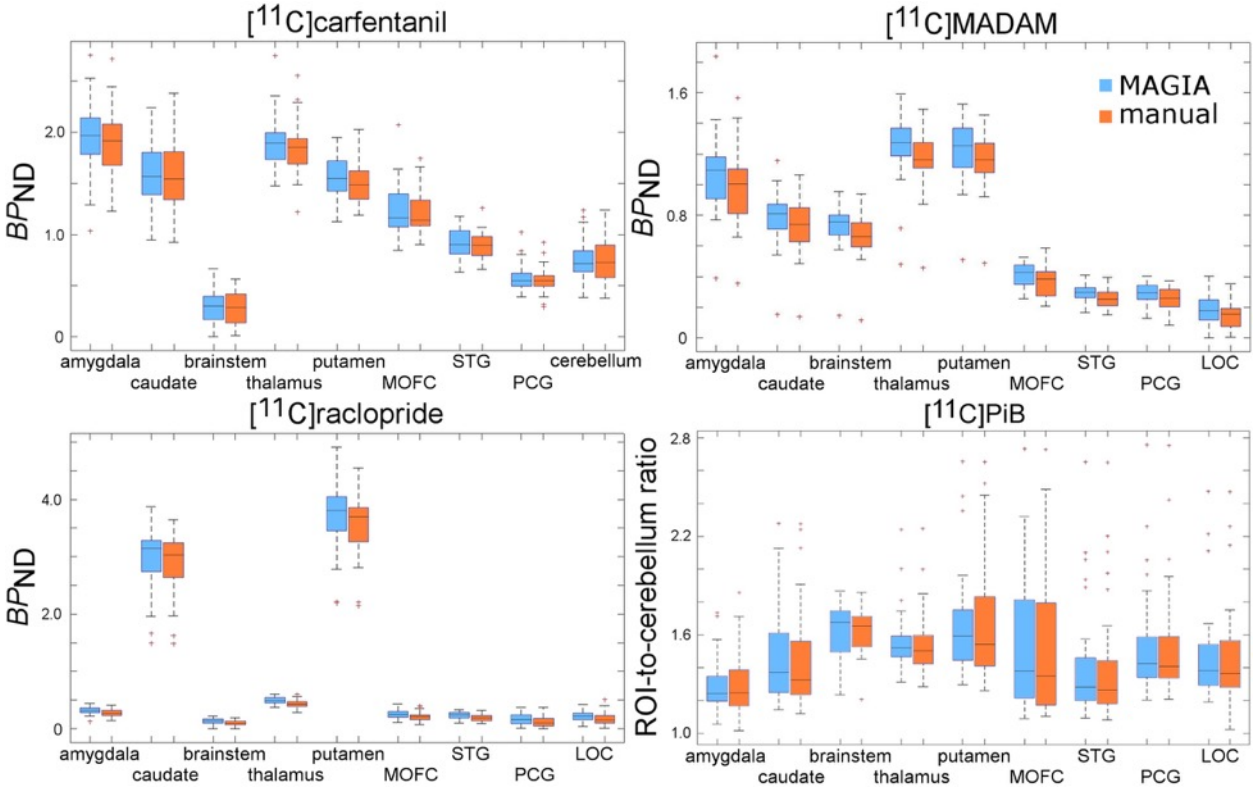

***Figure* 6.** *Boxplots of outcome measures in regions of interest derived from both automatic and manual reference regions. MOFC = medial orbitofrontal cortex, STG = superior temporal gyrus, PCG = postcentral gyrus, LOC = lateral occipital cortex.*

	[11C]raclopride					[11C]MADAM				
	BP_{ND} MAGIA	BP_{nd} manual	*p* -value	Diff%	Q1% - Q3%	BP_{ND} MAGIA	BP_{nd} manual	*p* -value	Diff%	Q1% - Q3%
amygdala	0.32	0.27	< 0.001	12.1	8.0 - 24.3	1.09	1.00	< 0.001	9.9	5.5 - 55.0
caudate	3.15	3.03	< 0.001	4.3	2.1 - 7.0	0.81	0.74	< 0.001	10.0	2.4 - 12.4
brainstem	0.14	0.09	< 0.001	40.0	16.6 - 53.2	0.75	0.66	< 0.001	13.9	3.9 - 20.1
thalamus	0.49	0.44	< 0.001	9.6	5.9 - 17.9	1.27	1.16	< 0.001	9.3	2.2 - 13.0
putamen	3.80	3.70	< 0.001	4.0	1.9 - 6.6	1.26	1.16	< 0.001	7.8	2.2 - 10.8
MOFC	0.26	0.21	< 0.001	17.9	6.3 - 33.2	0.43	0.38	< 0.001	16.8	4.6 - 27.0
STG	0.24	0.19	< 0.001	22.7	9.6 - 31.8	0.30	0.25	< 0.001	23.6	6.6 - 29.0
PCG	0.16	0.10	< 0.001	39.6	8.8 - 83.3	0.29	0.26	< 0.001	20.8	7.4 - 28.7
LOC	0.22	0.15	< 0.001	24.2	2.8 - 57.6	0.18	0.15	< 0.001	26.5	10.3 - 40.5

MOFC = medial orbitofrontal cortex, STG = superior temporal gyrus, PCG = postcentral gyrus, LOC = lateral occipital cortex

Table 2. Regional uptake estimates for [11C]raclopride and [11C]MADAM.

Figure 7 shows variability in the uptake estimates for one representative ROI per tracer (putamen for [11C]raclopride and [11C]PiB, and thalamus for [11C]MADAM and [11C]carfentanil). For each tracer, the manual estimates are shown in grey, while the Magia-derived estimates are shown in red. To aid visualization, between-study variability was removed by centering the uptake estimates for

each study separately. For [11C]PiB, Magia estimated the SUVR of one study to be more than two standard deviations away from the mean, while there were seven such outliers from the manual reference regions. For [11C]carfentanil, Magia did not produce any estimates outside the bounds defined by the two standard deviations. For [11C]raclopride, the Magia-derived estimates were consistently above means of the manual estimates, and 12 times above the upper bound, while there were five such manual estimates. For [11C]raclopride, in 12 cases Magia produced binding potential estimates at least two standard deviations greater than the mean of the manual estimates. There were nine manual binding potential estimates outside the bounds. Magia produced one estimate more than two standard deviations below the manual estimates for [11C]MAGIA, while there were seven outliers with the manual method.

The standard deviations of the regional uptakes for each tracer are also shown in Figure in the original uptake units. For Gaussian distributions, a range of two standard deviations symmetrically around the mean contains approximately 68 %, while four standard deviations cover already 95 % of the probability density. Thus, the 68 % and 95 % confidence intervals would span, in high-binding regions, approximately 0.2 and 0.4 SUVR-units for [11C]PiB, 0.5 and 0.9 for [11C]carfentanil BPND , 0.2 and 0.4 [11C]raclopride BPND, and 0.2 and 0.5 [11C]MADAM BPND. This uncertainty would arise only from subjective decisions related to delineation of reference regions.

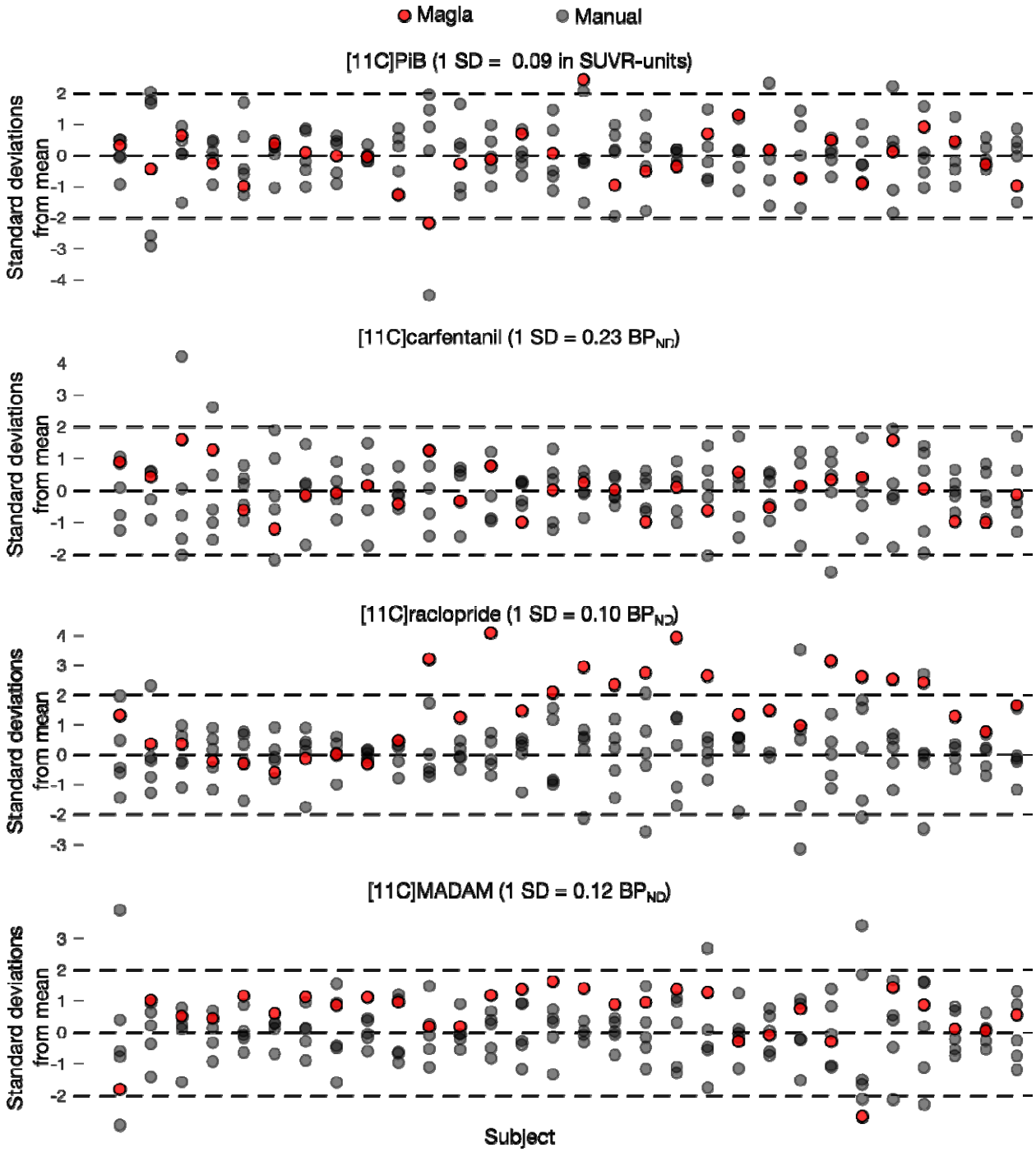

Figure 7. *Between-operator variance in representative ROIs. The horizontal lines reflect two standard deviations. Individual subjects are plotted along the x-axis.*

Discussion

We established that the fully automatic Magia pipeline yields consistent estimates of radiotracer uptake for all the tested ligands, with very little to no bias in the outcome measures. As expected, the manual delineation method suffered from significant operator-dependent variability, highlighting the importance of standardization of the process. This consistency coupled with significant gains in processing speed suggests that Magia is well suited for automated analysis of brain-PET data for large-scale neuroimaging projects.

Reliability of Magia's uptake estimates

Compared to averaged manual estimates, Magia produced parameter estimates without systematic bias for [11C]PiB SUVR and [11C]carfentanil BP. For [11C]PiB, the difference between the manual and automatic SUVR estimates fluctuated randomly around zero. Because SUVR was used to quantify [11C]PiB uptake, the random fluctuation was independent of brain region. For [11C]carfentanil, the random fluctuation was slightly greater in low-binding regions (but still within +/- 5 %). In contrast to [11C]PiB and [11C]carfentanil, there were systematic differences between the manual and automatic binding potential estimates for [11C]raclopride and [11C]MADAM. For both tracers the bias decreased as a function of specific binding, and in high-binding regions (BPND > 1.5) the bias was less than 5 %. Even if the bias increased sharply with decreasing binding potential, the problematic regions are not typically considered very interesting because of their poor signal-to-noise ratio.

The systematic bias for [11C]MADAM and [11C]raclopride is also reflected in the small differences in reference tissue TACs. For every cerebellar reference region, Magia-derived reference tissue TACs had 2-3 % lower AUCs. The peaks of the TACs were also slightly lower. For [11C]PiB, the bias did not propagate into outcome measures because the SUV-ratio was calculated between 60 and 90 minutes when there was no bias in TACs. Because binding potential reflects the ratio between specific binding and reference tissue signal, the ref TAC AUCs directly propagate into biases in binding potentials. Thus, these data indicate that Magia may produce slightly higher binding potential estimates than traditional methods if cerebellum is used as the reference region.

These data do not imply that the bias should be regarded as error: In fact, Magia produces significantly larger reference regions, and consequently the reference tissue TACs are less noisy. This is desirable, because the noise in input function influences model fitting. Despite this, the bias means that Magia-produced estimates should not be combined with estimates produced with other

methods. If all data are processed with Magia, however, there are no problems, because bias does not influence many population level analyses, such as between-subject correlation or group-difference analyses.

Variability in manual estimates

The present data highlight the importance of standardized definition of reference region definition. For all tracers, a substantial number of operator-wise estimates were at least 1 SD away from the mean of the estimates. The standard deviations were 0.1–0.2 in SUVR and BPND units. Thus, in the present study, it was not uncommon that differences between two outcome measure estimates derived by two individuals differed by more than two SD. Hence, even if the operators delineating the reference region had written instructions with pictures to help them, their outcome measure estimates often differed by 10–20 %. Magia generates reference regions using a standardized algorithm, thus substantially decreasing undesired variance in parameter estimates.

The automatic and manual reference regions differed in their topography. First, the automatic reference regions were consistently larger than their manually delineated counterparts. Only four studies had a smaller manual occipital cortex compared to their automatic counterparts. This was however expected as reference regions were drawn manually to only three transaxial slices, whereas FreeSurfer-defined region originally covered the whole region (either occipital cortex or cerebellum) which was subsequently trimmed down (see **Figure 2**). Manual delineation is typically limited to few slices because it is so labour intensive. Because increasing the number of voxels improves signal-to-noise ratio (REF), TACs based on larger ROIs are more reliable as long as the ROI is adequately placed. This latter aspect has however been well established for the FreeSurfer parcellations (Fischl et al., 2002). Second, there was surprisingly little overlap between the manual and automatic reference regions, as well as between the manually delineated ROIs within a subject. Poor overlap between manual and automatic reference regions is partly due to differences of their sizes, because the current standard for manual delineation involved only three transaxial slices. Additionally, FreeSurfer-based automatic reference regions follow strictly the cortical grey matter surface whereas manual reference regions may contain significant amounts of white matter because of the given instructions of reference region delineation in transaxial layer. Avoiding the white matter completely would further increase the time required foe manual ROI delineation. Finally, operators generating

the manual reference regions often chose different transaxial slices to draw the reference region, explaining most of the within-study anatomical differences in manual reference regions.

Functional homogeneousness of the reference regions

We tested whether the assumption of homogenous binding within the reference regions holds for both automatic and manual reference regions. A homogenous source region should produce unimodal and approximately symmetric radioactivity distributions (Teymurazyan et al., 2013). Between-study average distributions were unimodal and symmetric for all tracers for both the manual and automatic method. The distribution means were practically identical, but the modes were 1–2 % higher for Magia. The manual distributions were slightly wider (the standard deviations were approximately 15 % larger). Because Magia cuts the distribution tails, this was expected. The manual distributions were also slightly less skewed. Because averaging distributions tends to make them more Gaussian, this difference probably arises from the fact that the manual distributions that were used in the comparison were defined as an average over the individual manual distributions. The distribution overlaps were approximately 90 % for all tracers. In sum, these results show that the Magia-generated reference region radioactivity distributions are highly similar to the manually obtained distributions.

Reference tissue time-activity curves

Despite their topographical differences, the automatic and manual reference regions provided nearly identical time-activity curves. For all tracers, the Pearson correlation coefficient between average automatic and manual reference tissue TACs was above 0.99. This shows that the shapes of the TACs are almost identical. However, the AUCs of cerebellar time-activity curves were lower for Magia, indicating that the cerebellar automatic TACs were slightly positively biased compared to their manual counterparts.

Solving temporal constraints in processing of PET data

On average, drawing the reference region for one single study took around fifteen minutes if done carefully, and without any automatization the modeling and spatial processing of the images standard tools (e.g. PMOD or Turku PET Centre modelling software) takes easily at least 45 minutes. In contrast, Magia pipeline can be set running in less than five minutes per study. Although the time advantage–roughly an hour per study–gained from automatization is still modest in small-

scale studies (e.g. three eight-hour working days for a study with 24 subjects) the effect scales up quickly, and manual modeling of a database of just 400 studies would take already fifty days. This is significant investment of human resources, in particular if the analyses have to be redone later with, for example, different modeling parameters requiring repeating of at least some parts of the process.

Standardization of analysis methods

Functional neuroimaging community has already established standardized analysis pipelines for preprocessing fMRI data. However, a publicly available pipeline that automatically produces the outcome measures from PET images in a standardized fashion has been lacking. Of course, also the brain PET community has used standardized methods as much as possible. Magia only takes the standardization to extreme by providing a fully automated and standardized analysis option for brain PET studies. The increased standardization decreases variance resulting from subjective choices in the analysis process, thus improving estimation accuracy in population level analyses.

Magia is currently fully automatic only for studies for which a reference region exists. Thus, if plasma input function is needed, such as for Patlak or FUR, it needs to be produced before use in Magia. After that, Magia can handle processing automatically. Magia was originally developed with the assumption that T1-weighted MRI is available for each subject (for reference region delineation and normalization). Because this assumption limited the applicability of the approach for reanalysis of some historical data, Magia can now also use templates for ROI definition and tracer-specific radioactivity templates for spatial normalization. Thus, availability of MRI is no more a necessity, but it is recommended because most of the testing has been done with the MRI-based processing, and because the ROIs as well as reference regions are then generated in the native space.

Conclusions

Here we confirm that Magia is a standardized and fully automatic analysis pipeline for processing brain PET studies. By standardizing the reference region generation process, Magia removes substantial amount of variance in uptake estimates. For [11C]carfentanil that uses occipital cortex as the reference region, the reduced variance comes with no cost for bias in BPND. The SUVR estimates were also unbiased for [11C]PiB. [11C]raclopride and [11C]MADAM BPNDs were slightly overestimated. However, compared to the variance resulting from operator dependency, this bias was negligible, and in any case, it is meaningless in most population level analyses. Magia provides a

novel opportunity to reliably process large chunks of brain PET data, facilitating studies with large sample size.

References

Ashburner, J., et al. (2000). Voxel-Based Morphometry - the Methods. *Neuroimage, 11*, 805-821.

Button, K. S., et al. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience, 14*, 365-376.

Cremers, H. R., et al. (2017). The Relation between Statistical Power and Inference in Fmri. *PLoS One, 12*, 20.

Eickhoff, S. B., et al. (2005). A New Spm Toolbox for Combining Probabilistic Cytoarchitectonic Maps and Functional Imaging Data. *Neuroimage, 25*, 1325-1335.

Endres, C. J., et al. (2003). Quantification of Brain Mu-Opioid Receptors with C-11 Carfentanil: Reference-Tissue Methods. *Nucl Med Biol, 30*, 177-186.

Esteban, O., et al. (2019). Fmriprep: A Robust Preprocessing Pipeline for Functional Mri. *Nat Methods, 16*, 111-116.

Fischl, B., et al. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron, 33*, 341-355.

Gunn, R. N., et al. (1997). Parametric Imaging of Ligand-Receptor Binding in Pet Using a Simplified Reference Region Model. *Neuroimage, 6*, 279-287.

Lopresti, B. J., et al. (2005). Simplified Quantification of Pittsburgh Compound B Amyloid Imaging Pet Studies: A Comparative Analysis. *J Nucl Med, 46*, 1959-1972.

Lundberg, J., et al. (2005). Quantification of C-11-Madam Binding to the Serotonin Transporter in the Human Brain. *J Nucl Med, 46*, 1505-1515.

Poldrack, R. A., et al. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. In S. T. Fiske (Ed.), *Annual Review of Psychology, Vol 67* (Vol. 67, pp. 587-612). Palo Alto: Annual Reviews.

Schain, M., et al. (2014). Evaluation of Two Automated Methods for Pet Region of Interest Analysis. *Neuroinformatics, 12*, 551-562.

Simmons, J. P., et al. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*, 1359-1366.

Teymurazyan, A., et al. (2013). Properties of Noise in Positron Emission Tomography Images Reconstructed with Filtered-Backprojection and Row-Action Maximum Likelihood Algorithm. *Journal of Digital Imaging, 26*, 447-456.

Tuszynski, T., et al. (2016). Evaluation of Software Tools for Automated Identification of Neuroanatomical Structures in Quantitative Beta-Amyloid Pet Imaging to Diagnose Alzheimer's Disease. *European journal of nuclear medicine and molecular imaging, 43*, 1077-1087.

Tzourio-Mazoyer, N., et al. (2002). Automated Anatomical Labeling of Activations in Spm Using a Macroscopic Anatomical Parcellation of the Mni Mri Single-Subject Brain. *Neuroimage, 15*, 273-289.

White, D. R. R., et al. (1999). Intra- and Interoperator Variations in Region-of-Interest Drawing and Their Effect on the Measurement of Glomerular Filtration Rates. *Clin Nucl Med, 24*, 177-181.

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated Fmri Correlations Reflect Low Statistical Power-Commentary on Vul Et Al. (2009). *Perspect Psychol Sci, 4*, 294-298.

Yarkoni, T., et al. (2011). Neurosynth: A New Platform for Large-Scale Automated Synthesis of Human Functional Neuroimaging Data. *Frontiers in Neuroinformatics*.

Yasuno, F., et al. (2002). Template-Based Method for Multiple Volumes of Interest of Human Brain Pet Images. *Neuroimage, 16*, 577-586.