Severi Santavirta

MAGIA: ROBUST AUTOMATED MODELING AND IMAGE PROCESSING PLATFORM FOR PET NEUROINFORMATICS

Syventävien opintojen kirjallinen työ

Syyslukukausi 2018

Severi Santavirta

MAGIA: ROBUST AUTOMATED MODELING AND IMAGE PROCESSING PLATFORM FOR PET NEUROINFORMATICS

Turku PET Centre

Syyslukukausi 2018

Vastuuhenkilö: Prof. Lauri Nummenma

TURUN YLIOPISTO
Lääketieteellinen tiedekunta

SANTAVIRTA, SEVERI: Magia: robust automated modeling and image processing platform for PET neuroinformatics

Syventävien opintojen kirjallinen työ, 24 s.
Turku PET Centre
Lokakuu 2018

Aivojen neurobiologiaa voidaan molekyylitasolla tutkia PET-kuvantamisen avulla. Ennen statistisia analyysejä PET-data vaatii prosessoimista ja mallintamista. Alkuprosessointiin sisältyy useita vaiheita, kuten PET-kuvien liikekorjaus, kohdistaminen MRI-kuvan kanssa, kineettinen mallinnus sekä normalisaatio ja graafinen tasoitus (*smoothing*). Prosessointi- ja mallinnusvaiheet ovat usein erillisiä ja vaativat runsaasti aikaa. Perinteisesti kineettiseen mallinnukseen tarvittavat vertailualueet piirretään tutkittavien MRI-kuviin manuaalisesti, mikä vaatii tutkijalta huomattavasti työtä. Vaiheiden nopeuttamiseksi kehitimme Magian, joka on täysin automaattinen PET-kuvien prosessointi- ja mallinnustyökalu. Magia yhdistää jo olemassa olevia menetelmiä sekä uuden automaattisen menetelmän tuottaa vertailualueet PET-kuville.

Tässä tutkimuksessa validoimme Magian automaattista menetelmää tuottaa vertailualueet neljällä PET-merkkiaineella: [^{11}C]carfentanil, [^{11}C]raclopride, [^{11}C]MADAM ja [^{11}C]PiB. Valitsimme aiemmista tutkimuksistamme jokaiselle merkkiaineelle 30 tutkittavaa. Viisi aivotutkijaa piirsi manuaalisesti vertailualueet tutkittaville. Tämän jälkeen Magia-työkalun tuottamia vertailualueita verrattiin manuaalisiin vertailualueisiin. Tärkeimpänä automaattisen menetelmän luotettavuuden mittarina tutkimme menetelmien välisiä eroja merkkiaineiden sitoutumista kuvaavissa suureissa. BP_{ND}-arvoa (*binding potential*) käytettiin kuvaamaan [^{11}C]carfentanil-, [^{11}C]raclopride- ja [^{11}C]MADAM-merkkiaineiden sitoutumista. PiB-tutkimuksissa sitoutumista määritettiin SUVR-arvolla (*standardized uptake value ratio)*.

Merkittäviä eroja BP_{ND}-arvoissa [^{11}C]carfentanil-merkkiaineella ja SUVR-arvoissa [^{11}C]PiB-merkkiaineella ei todettu. [^{11}C]MADAM- ja [^{11}C]raclopride-merkkiaineilla automaattinen menetelmä tuotti merkitsevästi manuaalista suurempia BP_{ND}-arvoja. Korkean sitoutumisen alueilla ($BP_{ND} > 1$) BP_{ND}-arvojen ero oli korkeintaan 10 % ja matalan sitoutumisen alueilla ($BP_{ND} < 1$) ero vaihteli 17 %:n ja 40 %:n välillä.

Merkittäviä menetelmien välisiä eroja [^{11}C]carfentanil- ja [^{11}C]PiB-merkkiaineiden sitoutumisessa ei todettu. Magia tuotti [^{11}C]MADAM- ja [^{11}C]raclopride-merkkiaineilla systemaattisesti manuaalista menetelmää suurempia BP_{ND}-arvoja. PET-tutkimusten kannalta kiinnostavia ovat korkean sitoutumisen alueet, joissa todettua korkeintaan 10 %:n eroa voidaan pitää hyväksyttävänä. Todennäköisesti ero selittyy sillä, että Magian tuottamalla vertailualueella on vähemmän merkkiaineen sitoutumista reseptoriinsa. Löydösten perusteella Magian automaattinen vertailualueiden määritysmenetelmä on tutkituilla merkkiaineilla käyttökelpoinen.

Avainsanat: PET-tutkimus, mallinnus, prosessointi

TABLE OF CONTENTS

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

Tomi Karjalainen[1*], **Severi Santavirta[1]**, Tatu Kantonen[1], Jouni Tuisku[1], Lauri Tuominen[1,2], Jussi Hirvonen[1,3] 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

Keywords: PET, modelling, image processing, neuroinformatics, automatic reference region generation

Acknowledgements: This work was supported by the Academy of Finland grants #265915 and #294897 to LN and Sigrid Juselius Foundation grant to LN.

1 INTRODUCTION

Several publications have recently questioned statistical power of many neuroimaging studies. (Button et al., 2013). A shared conclusion of these publications is that larger sample sizes are needed. Simultaneously, 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, Nelson, & Simonsohn, 2011). Consequently, the focus in neuroimaging has shifted towards standardized, large-scale neuroinformatics based approaches (Poldrack & Yarkoni, 2016; Yarkoni, Poldrack, & Nichols, 2011). Today, several standardized and highly automatized preprocessing pipelines are publicly available for processing functional magnetic resonance images. Such standardized methods are not, however, currently available for analysis of positron emission tomography (PET) data.

The primary bottleneck for automatization of PET analysis is the requirement of input function. Depending on the 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. Optimally, the reference region should be defined separately for each individual before spatial normalization. Thus, the gold 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, Houston, Sampson, & Wilkins, 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 significantly larger number of scans should be processed.

To resolve these problems, we have introduced the Magia analysis pipeline for brain-PET data that enables automatic modeling 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 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.
5) Similarly with resting state fMRI pipelines, Magia produces the final first-level analysis results. This is in contrast with task fMRI studies in which statistical analysis depends on the task.

We verified the reliability of the automatic reference region generation, input function extraction, modeling, and spatial preprocessing of PET data with four tracers with different binding sites: [^{11}C]raclopride, [^{11}C]carfentanil, [^{11}C]MADAM, and [^{11}C]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.

2 MATERIALS AND METHODS

2.1 Magia platfrom

Magia (https://github.com/tkkarjal/magia) is a fully automatic analysis platform 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. Combining Magia with a database facilitates large-scale PET analyses. However, Magia can also be installed and used without such database.

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. In Magia each tracer has its 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.

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 co-registration with the MRI. The MRI is run through FreeSurfer to provide an anatomical label

for each voxel. The MRI is also segmented into grey and white matter probability maps for spatial normalization. The anatomical parcellation provided by FreeSurfer is used 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 before obtaining the input-function for modeling; the corrections are meant to make the reference region generation robust for many scanners and individuals. The subsequently obtained parametric images are normalized and smoothed. In addition to the parametric images, Magia also calculates region of interest (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.

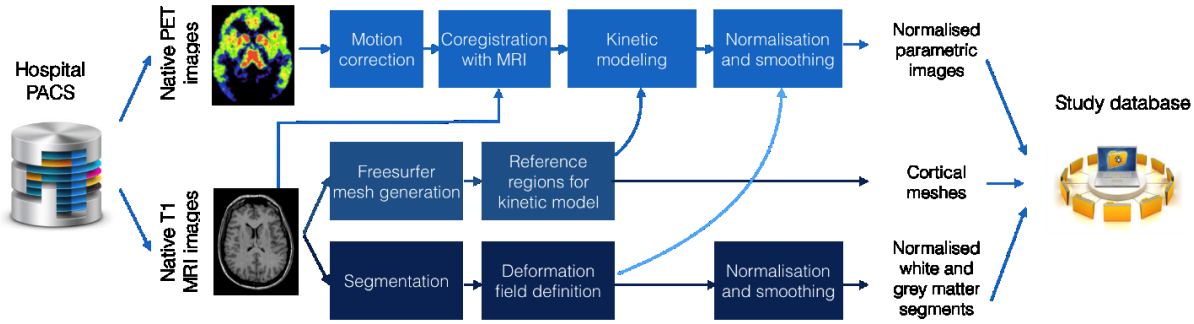

Figure 1. *The MAGIA pipeline combining FreeSurfer cortical mesh generation and parcellation, T1-weighted MRI image segmentation and normalization, automatic reference region and ROI generation, and kinetic modeling.*

All the steps mentioned above 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.

Magia requires MATLAB, SPM, and FreeSurfer and runs on Linux or Mac. The Optimization Toolbox for MATLAB is required for fitting the ROI level models. Magia has been developed using MATLAB R2016b.

2.2 Validation data

To assess reliability of Magia we used previously acquired data using four tracers binding to different binding sites: [^{11}C]raclopride, [^{11}C]carfentanil, [^{11}C]MADAM, and [^{11}C]PIB. For each tracer, we selected 30 studies from our previous experiments (Table 1). The validation focused on the reference region generation, because unlike other components of the pipeline, its reliability has not been previously tested. Thus, 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	GE Advance	HRRT	HRRT
	PET/CT	PET/CT		
	PET/MR	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).*

2.3 Manual reference region delineation

Five researchers with knowledge of human neuroanatomy delineated reference regions for every study according to written and visual instructions (Figure 1a). Cerebellum was used as a reference region for [^{11}C]raclopride (Gunn, Lammertsma, Hume, & Cunningham, 1997), [^{11}C]MADAM (Lundberg, Odano, Olsson, Halldin, & Farde, 2005) and [^{11}C]PiB (Lopresti et al., 2005). For [^{11}C]carfentanil, occipital cortex was used (Endres, Bencherif, Hilton, Madar, & Frost, 2003). The regions were drawn using CARIMAS (http://turkupetcentre.fi/carimas/).

The reference regions were defined on three consecutive transaxial T1-weighted MR images. 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 slice, 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.

2.4 Automatic reference region generation

Figure 2b shows an overview of the 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 have contained also specific binding. For cerebellum, the most important sources of spillover effects are occipital cortex and venous sinuses. Thus, the outermost cerebellar voxels are excluded in the anatomical reference region correction. For occipital cortex, voxels lateral to the lateral ventricles were excluded because the most lateral parts of the FreeSurfer-generated occipital cortex extend to areas with specific binding for [^{11}C]carfentanil. Also, the lateral ventricles provide an easy and reliable reference point for thresholding purposes. Finally, the radioactivity concentration distribution within the anatomically corrected reference region were estimated, and the tails of the distribution were excluded. The lower and upper boundaries for the signal intensities were defined by calculating the full width at half maximum (FWHM) of the mean PET signal intensity distribution and excluding voxels that were on the tail-ends of the corresponding radioactivity concentrations. This step ensured 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.

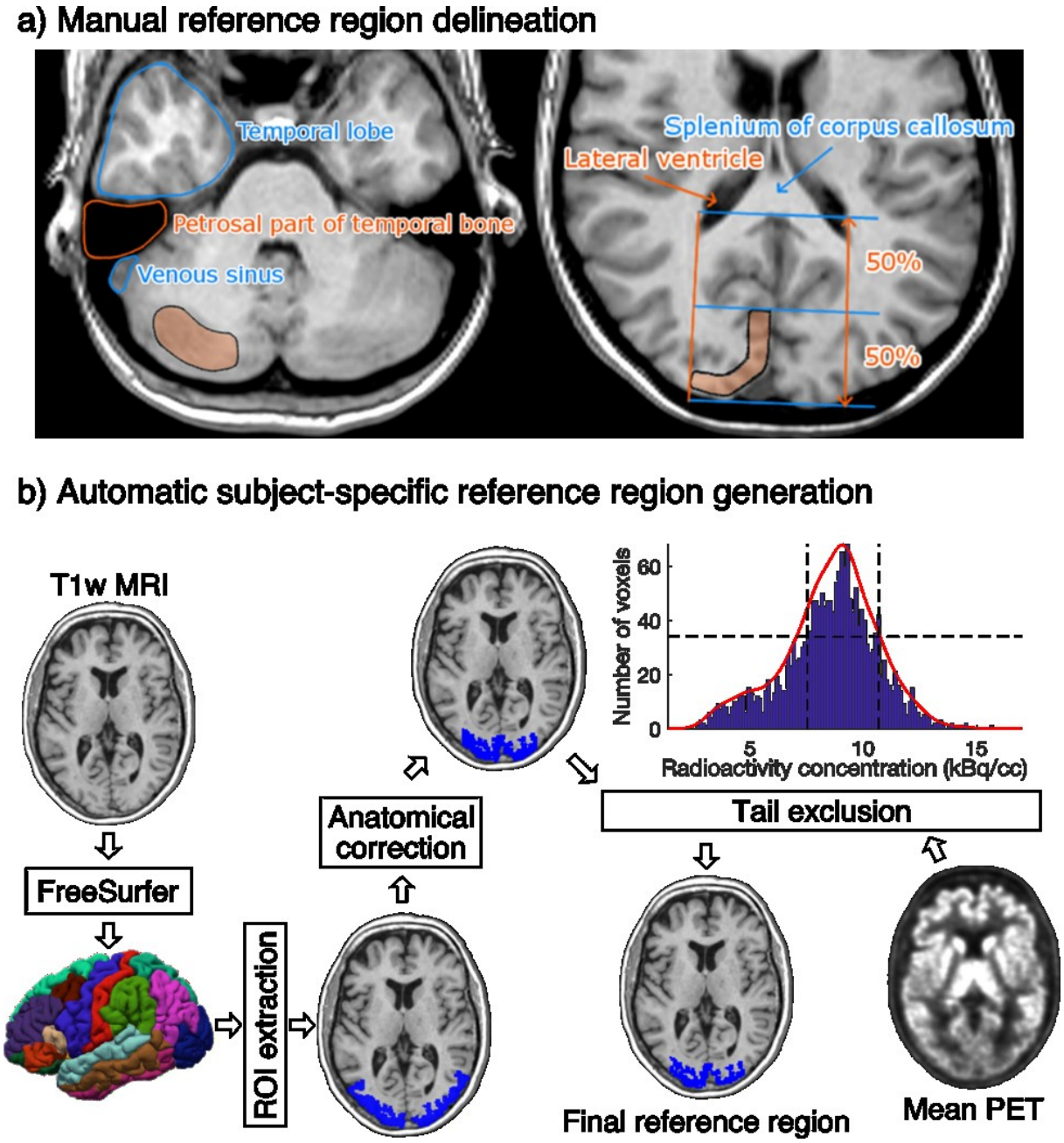

Figure 2. *a) Visual instructions of the most cranial slice of manually delineated cerebellar (left) and occipital (right) reference regions. The reference regions were delineated on three consecutive transaxial T1-weighted MR images. Cerebellar reference region is shown on the left and occipital reference 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 [^{11}C]carfentanil data set. The rectangles represent processing steps between inputs and outputs. The FreeSurfer step assigns an anatomical label for 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 [^{11}C]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.*

2.5 Validation metrics

2.5.1 Similarity of the uptake estimates

We used nondisplaceable binding potential (BP_{ND}) to quantify uptakes of [^{11}C]carfentanil, [^{11}C]raclopride and [^{11}C]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 [^{11}C]PiB uptake (Lopresti et al., 2005). All the studies were first processed using Magia and then 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 estimated the outcome measures in nine ROIs 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 [^{11}C]carfentanil and lateral occipital cortex (LOC) as a ROI for [^{11}C]raclopride, [^{11}C]PiB, and [^{11}C]MADAM. All ROIs were extracted from the FreeSurfer parcellations.

We also investigated 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.

2.5.2 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 manually delineated reference regions and compared it to the 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 the 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.

2.5.3 Similarity of the reference region radioactivity concentrations

A functionally homogenous region should have approximately Gaussian distribution of radioactivity measured with PET (Teymurazyan, Riauka, Jans, & Robinson, 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 distributions 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 SUVs, 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.

2.5.4 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).

2.5.5 Operator-dependent variability

We also 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 (BP_{ND}/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.

2.6 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).

3 RESULTS

3.1 Similarity of the uptake estimates

Figure 3 presents how the Magia-derived outcome measures differed from the average of the manual estimates in the full brain analysis. The average of manual estimates was regarded as the ground truth. For [^{11}C]MADAM, Magia produced up to 3–5 % higher binding potential estimates in regions with high specific binding. In cortical regions with low specific binding, the bias was over 10 %. For [^{11}C]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 [^{11}C]MADAM and [^{11}C]raclopride, the relative bias decreased significantly with increasing binding potential (Figure 3c). In contrast to these tracers, there was no systematic bias for [^{11}C]carfentanil or [^{11}C]PiB.

Figure 4 presents the results of outcome measures of each ROI for every tracer. In the ROI-based analysis, there also were no statistically significant differences of outcome measures in any ROI for [^{11}C]carfentanil and [^{11}C]PiB. However, significant differences were observed in every ROI for [^{11}C]raclopride and [^{11}C]MADAM. Magia produced up to 5 % higher BP_{ND} estimates for [^{11}C]raclopride in caudate and putamen which are well-known high-binding areas. Notably, estimates were significantly more variable in regions with no specific binding such as in cortex and brainstem (18 – 40 % higher with Magia), possibly reflecting increased signal in the larger ROIs in the areas containing mostly noise. Similarly, the bias in Magia produced BP_{ND} estimates for [^{11}C]MADAM were the lowest in high-binding areas (amygdala, thalamus, putamen) and the BP_{ND} difference was up to 10 % in these areas. The highest differences in

BP_{ND} estimates were observed in cortical low-binding areas (17 -27 %). Significant differences in outcome measures are shown in Table 2.

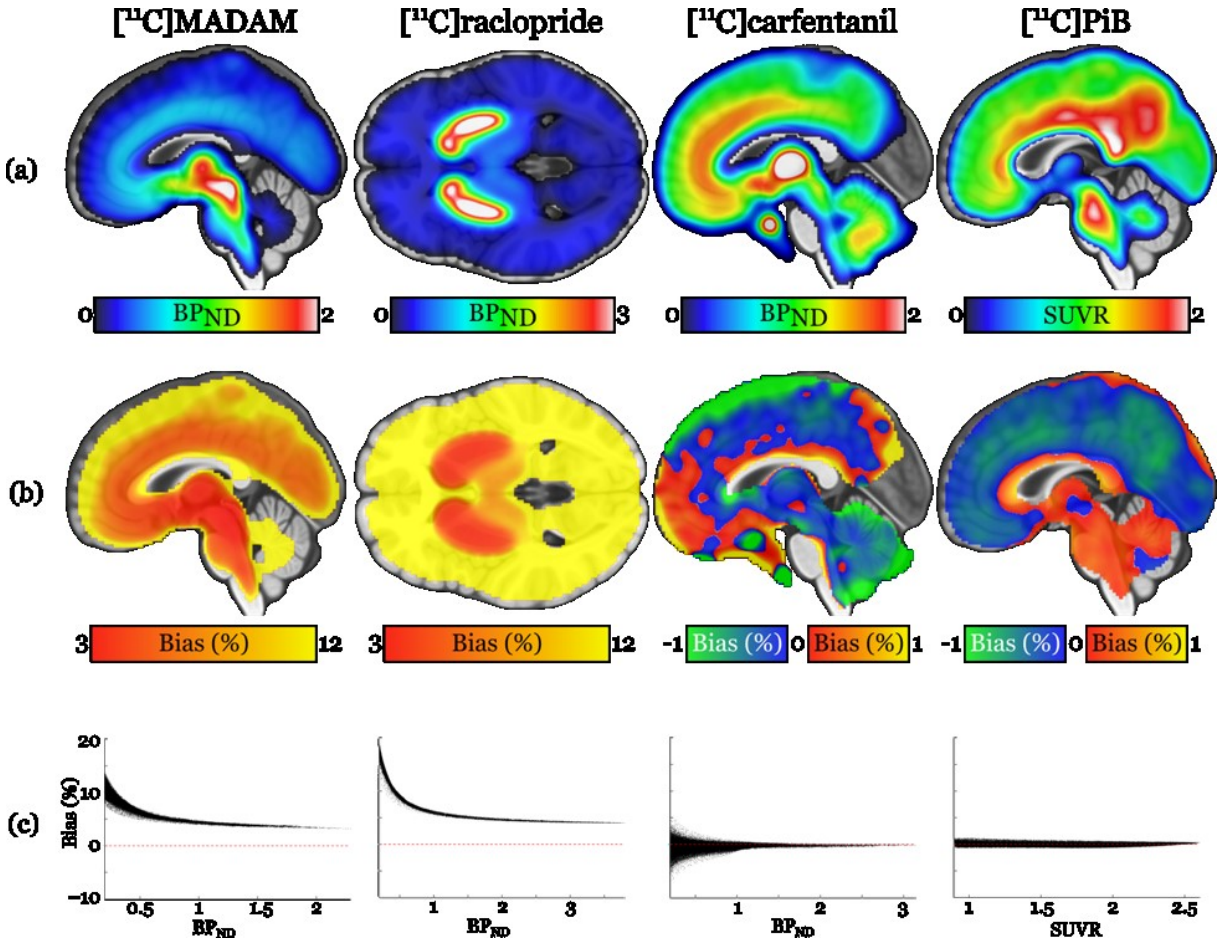

Figure 3. *(a) Visualization of the outcome measure distributions for each tracer. (b) Maps visualizing the relative biases of the Magia-derived outcome measures compared to the averages obtained by manual reference region delineation. The manual method is here presented as the ground truth, because 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.*

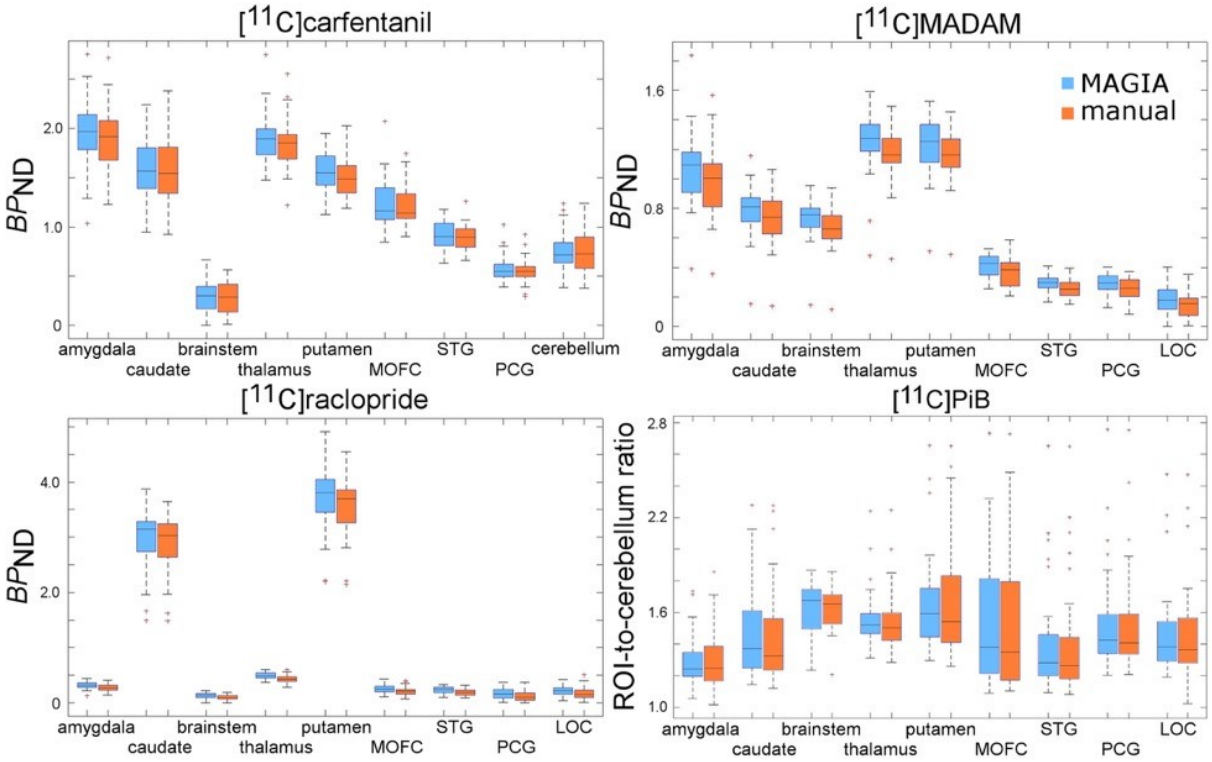

Figure 4. *Boxplots of outcome measures in regions of interest derived from both automatic and manual reference regions. MOFC = medial orbitofrontal cortex, STG = superior temporal gyrus, PCG = postcentral gyrus, LOC = lateral occipital cortex.*

	[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

Table 2. *Statistically significant differences in uptake estimates. MOFC = medial orbitofrontal cortex, STG = superior temporal gyrus, PCG = postcentral gyrus, LOC = lateral occipital cortex*

Figure 5 visualizes variability in the uptake estimates for one representative ROI per tracer. 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 [^{11}C]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 derived from the manual reference regions. For [^{11}C]carfentanil, Magia did not produce any estimates outside the bounds defined by the two standard deviations. For [^{11}C]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 [^{11}C]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 [^{11}C]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 5 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 [^{11}C]PiB, 0.5 and 0.9 for [^{11}C]carfentanil BP_{ND}, 0.2 and 0.4 [^{11}C]raclopride BP_{ND}, and 0.2 and 0.5 [^{11}C]MADAM BP_{ND}. This uncertainty would arise only from subjective decisions related to delineation of reference regions.

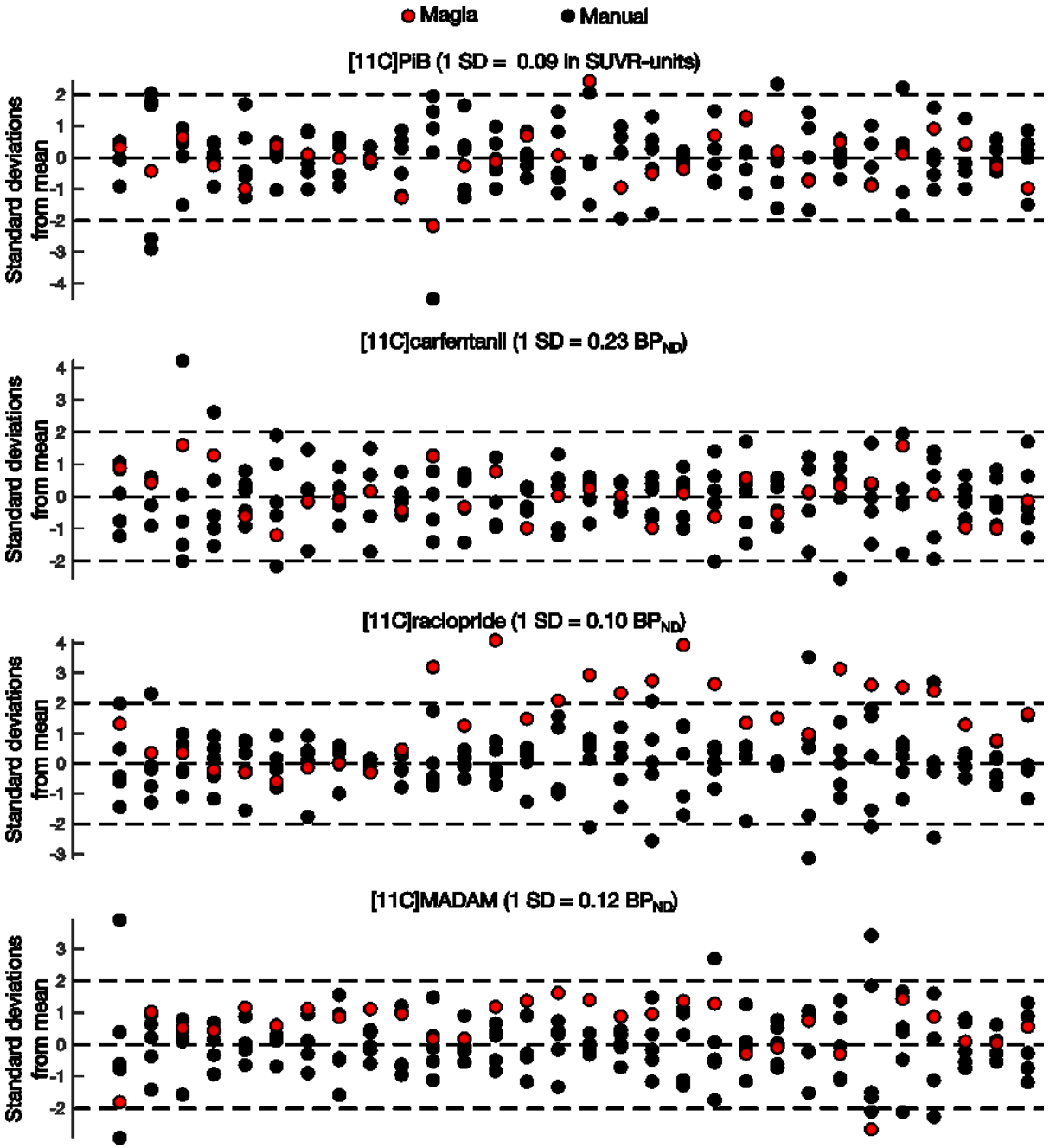

Figure 5. *Between-operator variance. The horizontal lines reflect two standard deviations.*

3.2 Functional properties of reference regions

3.2.1 Reference region SUV distributions

Mean reference region SUV distributions are shown in Figure 6a and time-activity curves of the reference regions in Figure 6b. 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 [^{11}C]carfentanil, [^{11}C]MADAM, [^{11}C]PIB and [^{11}C]raclopride, respectively. The modes of

the automatically and manually derived distributions were 1.5 and 1.55 for [^{11}C]carfentanil, 1.95 and 2.05 for [^{11}C]MADAM, 1.65 and 1.70 for [^{11}C]PIB, and 1.35 and 1.35 for [^{11}C]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 [^{11}C]carfentanil (24 % difference), 1.3 and 1.2 for [^{11}C]MADAM (11 % difference), 2.0 and 1.6 for [^{11}C]PIB (26 % difference), and 2.4 and 2.0 for [^{11}C]raclopride (21 % difference).

3.2.2 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 [^{11}C]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 [^{11}C]raclopride, [^{11}C]MADAM and [^{11}C]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 [^{11}C]carfentanil.

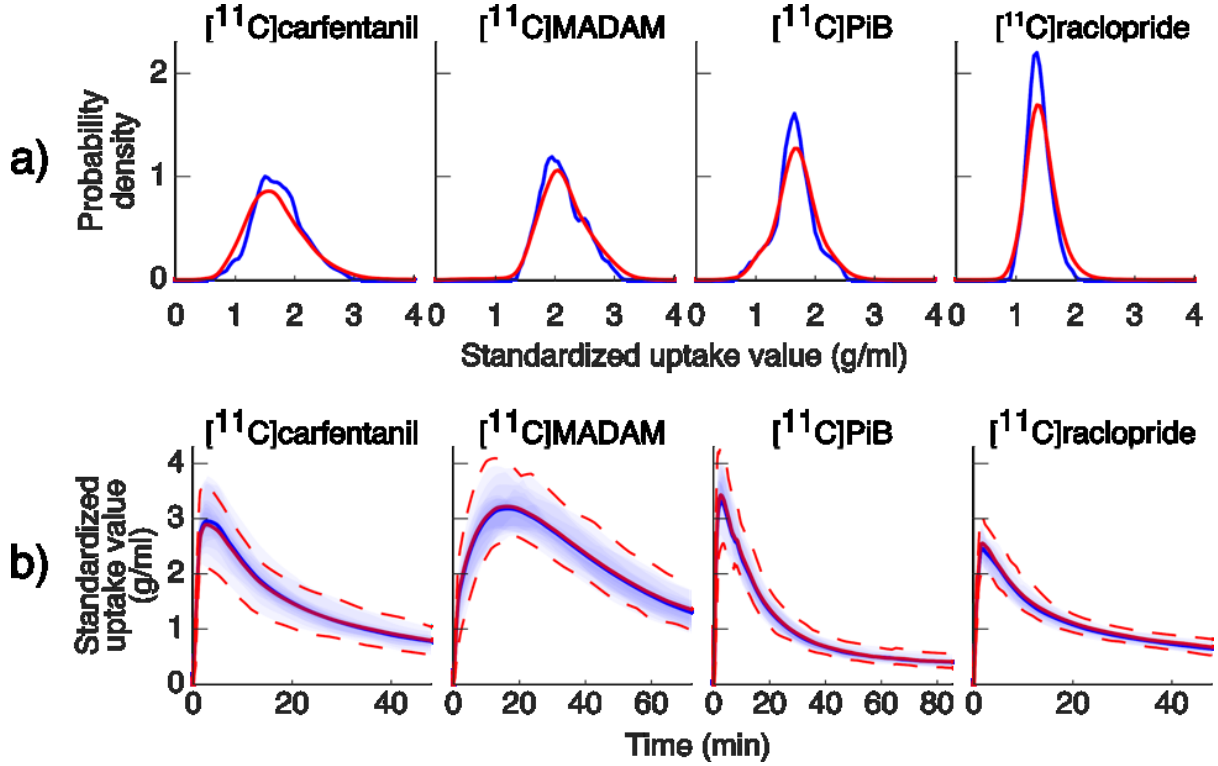

Figure 6. *a) Probability density distributions of the standardized uptake values within the reference regions. b) Automatic and manual reference region time-activity curves and the respective 80 % percentile intervals. Blue = Magia, red = manual.*

3.2.3 Within-study variation in manually obtained reference region 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 [^{11}C]carfentanil (12/20) occipital cortex was the reference region and the median difference of significant pairwise comparisons of AUCs was 9 %. For [^{11}C]raclopride (8/20), [^{11}C]MADAM (10/20) and [^{11}C]PiB (12/20) where cerebellum was the reference region median differences of significant comparisons of AUCs were 1 %, 2 %, 3 %, respectively.

3.3 Anatomical details of reference regions

3.3.1 Comparison of volumes between manual and automatic reference regions

For each tracer, automatic reference regions were, as expected, consistently larger than manually derived reference regions ($z > 4.35$, $p < 0.001$). The median ratios between volumes of automatic and manual reference regions were approximately 2 (Q1 - Q3: 1 - 2) for [^{11}C]carfentanil, 3 (Q1 - Q3: 2 - 4) for [^{11}C]raclopride, 8 (Q1 - Q3: 7 - 9) for [^{11}C]MADAM and 8 (Q1 - Q3: 7 - 9) for [^{11}C]PiB. Four [^{11}C]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. The volumes of reference regions are shown in Figure 7a).

3.3.2 Anatomical overlap between reference regions

We determined whether automatically determined reference regions overlap with the manually drawn reference regions. Automatic occipital reference region for [^{11}C]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 [^{11}C]raclopride, [^{11}C]MADAM and [^{11}C]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 2.4 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.

3.3.3 Topographical within-study variation in manual reference regions

Figure 7b shows a representative example of the topographical variation of the manual delineations of cerebellum and occipital cortex. Tracer level median overlaps between manual drawers were 22 %, 41 %, 14 %, 18 %, for [^{11}C]carfentanil, [^{11}C]raclopride, [^{11}C]MADAM and [^{11}C]PiB, respectively. Poor overlap can be mostly explained by the fact that drawers often chose different transaxial slices of MR images to draw the reference regions.

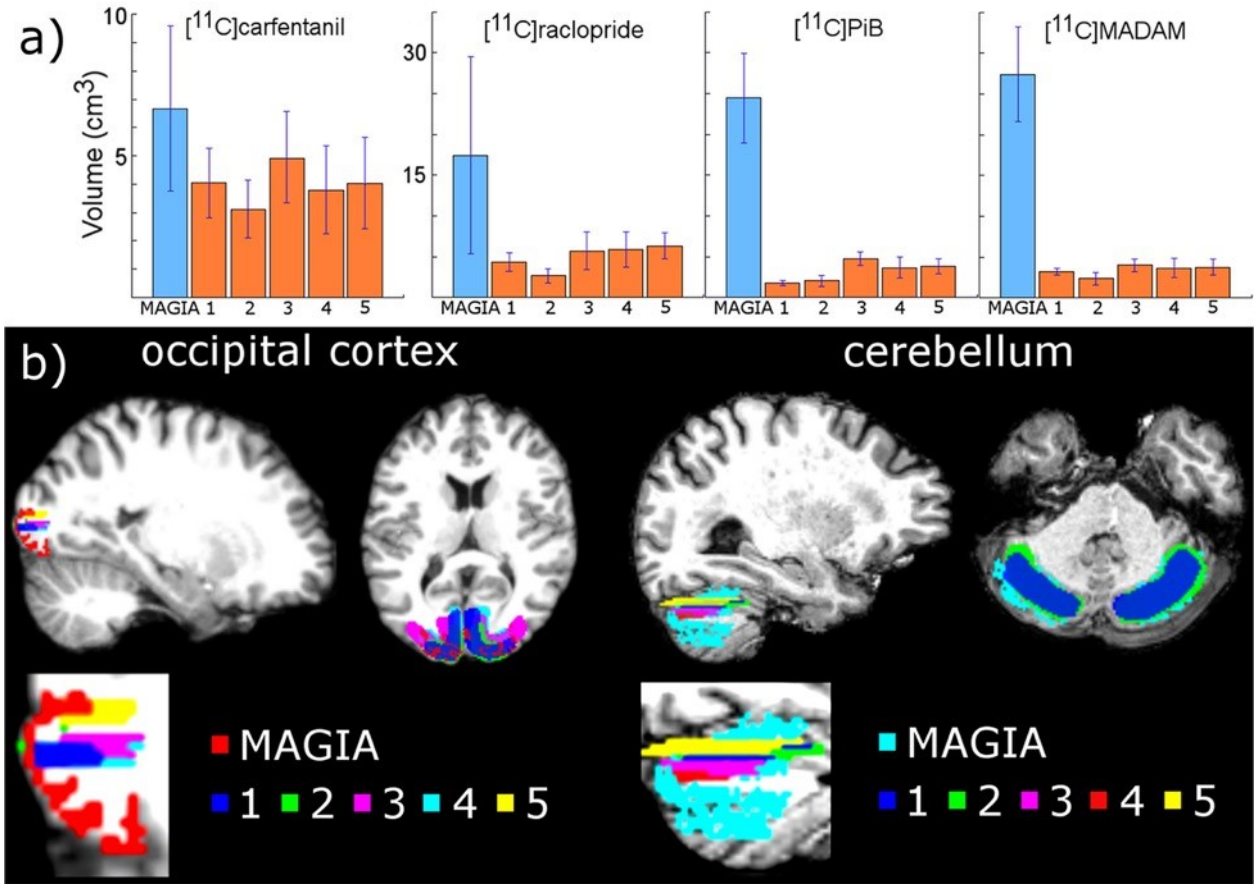

Figure 7. *a) Mean volumes of MAGIA-generated reference regions compared to mean volumes of manually delineated reference regions. b) Visual example of MAGIA-generated and manual reference regions for one study.*

4 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.

4.1 Reliability of Magia's uptake estimates

Compared to averaged manual estimates, Magia produced parameter estimates without systematic bias for [^{11}C]PiB SUVR and [^{11}C]carfentanil BP_{ND}. For [^{11}C]PiB, the difference between the manual and automatic SUVR estimates fluctuated randomly around zero. Because SUVR was used to quantify [^{11}C]PiB uptake, the random fluctuation was independent of brain region. For [^{11}C]carfentanil, the random fluctuation was slightly greater in low-binding regions (but still within +/- 5 %). In contrast to [^{11}C]PiB and [^{11}C]carfentanil, there were systematic differences between the manual and automatic binding potential estimates for [^{11}C]raclopride and [^{11}C]MADAM. For both tracers the bias decreased as a function of specific binding, and in high-binding regions ($BP_{ND} > 1.5$) 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 [^{11}C]MADAM and [^{11}C]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 [^{11}C]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 reference region 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 good because the noise in input function influences model fitting. Having said that, 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.

4.2 Variability in manual estimates

The present data illuminate the importance of highly standardized definition of reference region definition. For all tracers, a substantial number of subjective estimates were at least one SD away from the mean of the estimates. The standard deviations were 0.1 - 0.2 in SUVR and BP_{ND} 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. Thus, in the present study, even if the persons 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.

4.3 Reference region topography

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 1). 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, TACs based on larger ROIs are more reliable if the ROI is adequately placed. This latter aspect has however been well established for the FreeSurfer parcellations (Fischl B *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. 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. 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.

4.4 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 with the manually obtained distributions.

4.5 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 automatic and average manual reference tissue TAC 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.

4.6 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 with 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.

4.7 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.

4.8 Limitations

Magia does not work on Windows computers. 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 fully processed before use in Magia. Currently Magia recognizes only cerebellum or occipital cortex as reference regions; however, also other regions can be added if necessary. Finally, the present approach requires that T1-weighted MRI is available for each subject (for reference region delineation and normalization), limiting the applicability of the approach for re-analysis of some historical data.

5 CONCLUSIONS

Magia is a standardized and fully automatic analysis pipeline for processing brain PET studies and is publicly available in https://github.com/tkkarjal/magia. By standardizing the reference region generation process, Magia removes substantial amount of variance in uptake estimates. For [^{11}C]carfentanil that uses occipital cortex as the reference region, the reduced variance comes with no cost for bias in BP_{ND}. The SUVR estimates were also unbiased for [^{11}C]PiB. [^{11}C]raclopride and [^{11}C]MADAM BP_{ND}s 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 amounts of brain PET data, facilitating studies with large sample size.

REFERENCES

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034

Endres, C. J., Bencherif, B., Hilton, J., Madar, I., & Frost, J. J. (2003). Quantification of brain μ-opioid receptors with [11C]carfentanil: Reference-tissue methods. *Nuclear Medicine and Biology*, *30*(2), 177–186. https://doi.org/10.1016/S0969-8051(02)00411-0

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Gunn, R. N., Lammertsma, A. A., Hume, S. P., & Cunningham, V. J. (1997). Parametric imaging of ligand-receptor binding in PET using a simplified reference region model. *NeuroImage*, *6*(4), 279–287. https://doi.org/10.1006/nimg.1997.0303

Lopresti, B. J., Klunk, W. E., Mathis, C. A., Hoge, J. A., Ziolko, S. K., Lu, X., … Price, J. C. (2005). Simplifed Quantification of Pittsburgh Compound B Amyloid Imaging PET Studies: A Comparative Analysis. *Journal of Nuclear Medicine*, *46*, 1959–1972.

Lundberg, J., Odano, I., Olsson, H., Halldin, C., & Farde, L. (2005). Quantification of 11C-MADAM binding to the serotonin transporter in the human brain. *Journal of Nuclear Medicine*, *46*(9), 1505–1515. https://doi.org/46/9/1505 [pii]

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annu Rev Psychol.*, *67*, 587–612. https://doi.org/10.1016/S2214-109X(16)30265-0.Cost-effectiveness

Schain, M., Varnäs, K., Cselényi, Z., Halldin, C., Farde, L., & Varrone, A. (2014). Evaluation of Two Automated Methods for PET Region of Interest Analysis. *Neuroinformatics*, *12*(4), 551–562. https://doi.org/10.1007/s12021-014-9233-6

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Teymurazyan, A., Riauka, T., Jans, H. S., & Robinson, D. (2013). Properties of noise in positron emission tomography images reconstructed with filtered-backprojection and row-action maximum likelihood algorithm. *Journal of Digital Imaging*, *26*(3), 447–456. https://doi.org/10.1007/s10278-012-9511-5

Tuszynski, T., Rullmann, M., Luthardt, J., Butzke, D., Tiepolt, S., Gertz, H. J., … Barthel, H. (2016). Evaluation of software tools for automated identification of neuroanatomical structures in quantitative β-amyloid PET imaging to diagnose Alzheimer's disease. *European Journal of Nuclear Medicine and Molecular Imaging*, *43*(6), 1077–1087. https://doi.org/10.1007/s00259-015-3300-6

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

White, D. R. R., Houston, A. S., Sampson, W. F. D., & Wilkins, G. P. (1999). Intra- and interoperator variations in region-of-interest drawing and their effect on the measurement

of glomerular filtration rates. *Clinical Nuclear Medicine*, *24*(3), 177–181. https://doi.org/10.1097/00003072-199903000-00008

Yarkoni, T., Poldrack, R., & Nichols, T. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635.Large-scale

Yasuno, F., Hasnine, A. H., Suhara, T., Ichimiya, T., Sudo, Y., Inoue, M., … Toyama, H. (2002). Template-based method for multiple volumes of interest of human brain PET images. *NeuroImage*, *16*(3 I), 577–586. https://doi.org/10.1006/nimg.2002.1120