Citation for published version (APA):
Multi-atlas attenuation correction supports full quantification of static and dynamic brain PET data in PET-MR

To cite this article: Inés Mérida et al 2017 Phys. Med. Biol. 62 2834

View the article online for updates and enhancements.
Multi-atlas attenuation correction supports full quantification of static and dynamic brain PET data in PET-MR

Inés Mérida1,2,3, Anthonin Reilhac3, Jérôme Redouté3, Rolf A Heckemann5,6, Nicolas Costes3,8,9, Alexander Hammers4,7,8

1 Université de Lyon 1, INSERM, CNRS, Lyon Neuroscience Research Center, Lyon, France
2 Siemens Healthcare France SAS, Saint-Denis, France
3 CERMEP-Imagerie du vivant, Lyon, France
4 Neurodis Foundation, Lyon, France
5 MedTech West at Sahlgrenska University Hospital, Gothenburg, Sweden
6 Division of Brain Sciences, Imperial College London, London, United Kingdom
7 King’s College London & Guy’s and St Thomas’ PET Centre, Division of Imaging Sciences and Biomedical Engineering, Kings’ College London, London, United Kingdom

E-mail: costes@cermep.fr

Received 23 November 2016
Accepted for publication 9 February 2017
Published 14 March 2017

Abstract
In simultaneous PET-MR, attenuation maps are not directly available. Essential for absolute radioactivity quantification, they need to be derived from MR or PET data to correct for gamma photon attenuation by the imaged object. We evaluate a multi-atlas attenuation correction method for brain imaging (MaxProb) on static [18F]FDG PET and, for the first time, on dynamic PET, using the serotoninergic tracer [18F]MPPF.

A database of 40 MR/CT image pairs (atlases) was used. The MaxProb method synthesises subject-specific pseudo-CTs by registering each atlas to the target subject space. Atlas CT intensities are then fused via label propagation and majority voting. Here, we compared these pseudo-CTs with the real CTs in a leave-one-out design, contrasting the MaxProb approach...
with a simplified single-atlas method (*SingleAtlas*). We evaluated the impact of pseudo-CT accuracy on reconstructed PET images, compared to PET data reconstructed with real CT, at the regional and voxel levels for the following: radioactivity images; time-activity curves; and kinetic parameters (non-displaceable binding potential, BPND).

On static [18F]FDG, the mean bias for *MaxProb* ranged between 0 and 1% for 73 out of 84 regions assessed, and exceptionally peaked at 2.5% for only one region. Statistical parametric map analysis of *MaxProb*-corrected PET data showed significant differences in less than 0.02% of the brain volume, whereas *SingleAtlas*-corrected data showed significant differences in 20% of the brain volume. On dynamic [18F]MPPF, most regional errors on BPND ranged from -1 to +3% (maximum bias 5%) for the *MaxProb* method. With *SingleAtlas*, errors were larger and had higher variability in most regions. PET quantification bias increased over the duration of the dynamic scan for *SingleAtlas*, but not for *MaxProb*. We show that this effect is due to the interaction of the spatial tracer-distribution heterogeneity variation over time with the degree of accuracy of the attenuation maps.

This work demonstrates that inaccuracies in attenuation maps can induce bias in dynamic brain PET studies. Multi-atlas attenuation correction with *MaxProb* enables quantification on hybrid PET-MR scanners, eschewing the need for CT.

Keywords: magnetic resonance imaging, positron emission tomography, pseudo-CT, attenuation map, kinetic modelling

(Some figures may appear in colour only in the online journal)
optimization problem widen the set of possible solutions, and the algorithm may well converge on local minima leading to inconsistent emission and attenuation maps, leading to crosstalk artefacts. Recent variants use anatomical information derived from the subject MRI in order to guide the optimization process (Salomon et al 2011, Mehranian and Zaidi 2015). Those refined approaches produce encouraging results with static emission data but have so far been inferior in brain studies to multi-atlas approaches in direct comparisons (Mehranian et al 2016). They have never been assessed in dynamic brain imaging with changing counting statistics and activity distributions across time. As their performance depends on the tracer used, it is unlikely that they would be widely applicable in research centres using multiple tracers.

The second group of methods segments the subject’s MR images into material classes (mostly air, soft tissue and bone) and assigns to each of them a representative constant attenuation coefficient. (We follow the established terminology and refer to the classes as ‘tissue’ classes, even though air is not actually a tissue). Zaidi et al (2003) segment the T1-weighted MR image into four tissue classes (air, sinus, bone and tissue) using a fuzzy clustering technique. In contrast to T1-weighted MR images, the UTE sequence allows, to some extent, the distinction of bone signal from air. In Keereman et al (2010) it is used to obtain a more accurate bone segmentation. Poynton et al (2014) combine probabilistic segmentation of T1-weighted and UTE sequences with a probabilistic CT atlas producing an improved segmentation of the MR image into air, soft tissue, and bone. After segmentation, those methods generally assume a constant attenuation coefficient per tissue class. This may not be representative of the actual local tissue density, which can induce inaccuracies in reconstructed PET images. This is particularly true for osseous tissues that exhibit a large range of densities as shown by Catana et al (2010). Recently, Juttukonda et al (2015) and Ladefoged et al (2015) have proposed approaches that attempt to model the CT image intensity from the UTE signal for the bone class, allowing the computation of continuous attenuation coefficients for bone, using constant coefficients only for the remaining tissue classes. While these approaches have produced encouraging results, their accuracy still strongly relies on the exactness of the initial tissue classification.

The last family of methods uses a database of image pairs (MR and CT or MR and PET images) to derive a pseudo \( \mu \)-map. Unlike MR-based segmentation techniques, this process generates continuous attenuation coefficients for the whole volume. Some approaches utilize the database in a learning step to establish a model linking MR and CT intensities based on local image features and using either a Gaussian mixture regression model (Johansson et al 2011, Larsson et al 2013) or a super-vector regression model (Navalpakkam et al 2013). This model is then used to derive for each voxel of the subject MRI the corresponding CT intensity. Patch-based techniques use a database of coregistered MR and CT image pairs to predict a subject-specific pseudo-CT by performing an intensity-based nearest neighbour search between patches extracted from the subject MRI and patches extracted from a database (Torrado-Carvajal et al 2015, Andreasen et al 2016). Such approaches are promising but have not yet been evaluated exhaustively.

Using a multimodality optical flow deformable model, Schreibmann et al (2010) propose to create a simulated CT image that matches the patient anatomy by mapping the CT image of a single subject to the patient space. In other approaches, a single template is built by averaging several subjects from the database registered to a common space (Montandon and Zaidi 2005, Malone et al 2011, Izquierdo-Garcia et al 2014). This single template is then warped into the subject space with a single registration to derive an attenuation map that is subject-specific to varying degrees. Finally, true multi-atlas approaches have been used in the MRAC context to generate subject-specific pseudo-CTs from a database of MR and CT pairs (Burgos et al 2015, 2014a, Sjölund et al 2015, Mehranian et al 2016). In contrast to single atlas and
template methods, true multi-atlas methods register all CT-MR atlas pairs independently, thereby reducing the influence of errors in the individual registrations. To the extent that such errors are uncorrelated, they tend to cancel each other out. Multi-atlas techniques have been proposed originally for image segmentation problems, in particular for brain segmentation into anatomical regions where it has been shown that they outperform methods that use averaging prior to registration (i.e. template methods) by a large margin (Heckemann et al 2006). Independent registrations also give the opportunity to better address inter-subject variability by selecting, in the final step, the most relevant information from the database based on local features. Multi-atlas approaches outperform single atlas methods (Burgos et al 2014a) and template methods (Burgos et al 2015) for the generation of pseudo-CT images.

All proposed attenuation map generation methods have only been assessed in the context of static PET acquisition. Recent work (Ladefoged et al 2016) has shown that different AC methods perform differently depending on the PET tracer used ([18F]FDG, [11C]PiB and [18F] flurbiprofen). Those results suggest that the AC performance may depend on the tracer spatial distribution (variation of contrast) in brain. Research brain PET studies typically use dynamic acquisitions in which tracer spatial distribution changes over time, and no performance data on MR-based AC for dynamic PET imaging have been published so far. In this work we use dynamic PET data to explore this phenomenon.

In a recent study (Mérida et al 2015), we introduced a multi-atlas approach used to synthesise a subject-specific pseudo-CT by registering individual atlases to the target subject space and fusing atlas CT intensities via label propagation and majority voting (MaxProb method). Here, using an improved atlas database with more subjects and higher CT resolution, we provide a complete evaluation of the MaxProb method on quantitative static PET data, and also, for the first time, on dynamic PET data. Static evaluation was based on [18F] FDG PET as this tracer is widely used in clinical and research applications. Its homogeneous uptake in the whole brain allows a global evaluation of MRAC. Dynamic assessment was performed with 2′-methoxyphenyl-(N-2′-pyridinyl)-p-[18F]fluoro-benzamidoethylpiperazine ([18F]MPPF), a selective antagonist at 5-HT1A receptors found mainly in limbic structures. Evaluation includes accuracy on binding parameters estimated from kinetic modelling using a reference region.

1. Materials and methods

1.1. Materials

1.1.1. Atlas database. Data were available for 40 subjects (13 males, 27 females) (mean age ± SD, 33.9 ± 13.2 years; range, 16–63 years), selected as a convenience sample from our research database on the basis of their PET/CT and MR availability. Data used in this study are anonymized images of subjects who had participated in various ethically approved research studies. The anonymization procedure was registered under the number 1134516 by the competent authority (Comité de Protection des Personnes Sud-Est III). Subjects had been informed that their anonymized images could be used for methodological development, and had been given the option to oppose this use of their data. The subjects’ MR images were visually reviewed for conspicuous brain abnormalities (none found). Each subject had a T1-weighted MR image and a PET/CT brain scan. Three-dimensional anatomical T1-weighted sequences (MPRAGE) were acquired on a Siemens Sonata 1.5 Tesla MR scanner (TE = 3.93 ms, TR = 1970 ms, flip angle = 15°). The images were reconstructed in a 256 × 256 × 176 matrix with voxel dimensions of 1 × 1 × 1 mm3. CT images were acquired
on a Siemens Biograph mCT PET/CT tomograph at the energy of 80 keV. The images were reconstructed in a $512 \times 512 \times 149$ matrix with a voxel size of $0.58 \times 0.58 \times 1.5 \text{mm}^3$. MR images were corrected for field inhomogeneities using SPM12 (Statistical Parametric Mapping 12; Wellcome Trust Centre for Neuroimaging, UCL, London, UK). Each subject’s field-bias corrected MR image was aligned with the CT image using the affine registration tool reg_aladin from the NiftyReg software suite, optimizing normalized cross correlation for the image pair (http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg (Ourselin et al 2001)). Coregistered MR images were resampled to their initial resolution using cubic Hermite spline interpolation. Voxel values in CT images quantitatively represent radiodensity in Hounsfield units (HU). We therefore chose CT image as the reference space in order to avoid interpolation of these values. We use the term atlas to refer to the resulting CT and coregistered T1 MRI image pair.

1.1.2. Test data.

1.1.2.1. PET scanning. From the 40 subjects of the database, 30 had undergone a static $[^{18}\text{F}]$FDG or a dynamic $[^{18}\text{F}]$MPPF PET acquisition, and the corresponding data were used in the assessment. PET scans were obtained on the same Siemens Biograph mCT PET/CT tomograph as the CT scans. Twenty-three subjects (mean age $\pm$ SD, 35.0 $\pm$ 14.5 years; range, 16–63 years) had a 10-minute PET scan, obtained from 40 to 50 min after the injection of $125 \pm 26.4$ MBq of $[^{18}\text{F}]$FDG. Seven subjects (mean age $\pm$ SD, 33.4 $\pm$ 9.8 years; range, 19–44 years) had a dynamic PET scan during 60 min starting with the injection of $164 \pm 42.6$ MBq of $[^{18}\text{F}]$MPPF. All data were acquired in list mode.

1.1.2.2. PET reconstruction. PET data were reconstructed with an offline version of the Siemens reconstruction software (e7tools, Siemens Medical Solutions, Knoxville, USA). Actual CT images were converted to attenuation maps ($\mu$-map) by applying a bilinear transformation (Carney et al 2006) followed by Gaussian blurring (FWHM = 4 mm), and resampled to the PET voxel grid. $[^{18}\text{F}]$FDG data were rebinned into a single 10 min frame, whereas $[^{18}\text{F}]$MPPF data were rebinned into 35 time frames (variable length frames, $15 \times 20$ s, $15 \times 120$ s, $5 \times 300$ s) for dynamic reconstruction. Images were reconstructed using two different algorithms: 1) 3D ordinary Poisson-ordered subsets expectation maximization (OP-OSEM 3D) incorporating the system point spread function using 12 iterations of 21 subsets and 2) 2D Fourier rebinning (FORE) followed by 2D filtered-back projection (FBP2D) using a ramp filter with a cut-off at Nyquist frequency. Data correction (normalization, attenuation and scatter correction) occurred either before reconstruction (FBP2D) or was fully integrated within the reconstruction process (OP-OSEM 3D). Time-of-flight was not used, as the PET-MR Siemens Biograph mMR system for which the method is intended does not record time-of-flight information. Gaussian post-reconstruction filtering (FWHM = 4 mm) was applied to all PET images. Reconstructions were performed with a zoom of 3 yielding a voxel size of $1.06 \times 1.06 \times 2.02 \text{mm}^3$ in a matrix of $128 \times 128 \times 109$ voxels.

1.1.2.3. MRI segmentation. The T1 MR images were anatomically segmented into 83 regions using a maximum probability atlas in Montreal Neurological Institute (MNI)/International Consortium for Brain Mapping stereotaxic space, based on manual delineations of 30 MRIs of healthy young adults (Hammers_mith maximum probability atlas n30r83, (Hammers et al 2003, Gousias et al 2008), available at www.brain-development.org). An 84th region, the cerebellar vermis, was manually added in stereotaxic space. Deformation fields from the subjects’ space to MNI space were determined from the T1 MR image by using the Segment function of SPM12. The atlas was denormalized to each individual MRI space via the inverse transformation. Masks of grey matter (GM), white matter (WM), and cerebrospinal fluid
(CSF) were generated in the subject space by combining the Hammersmith MRI segmentations and the probabilistic ‘tissue’ maps obtained with SPM12 (SPM Segment).

1.1.2.4. Modelling of $[^{18}\text{F}]$MPPF. The $[^{18}\text{F}]$MPPF data were modelled with the simplified reference tissue model as described in Costes et al (2005). Parametric BPND images of 5-HT$_{1A}$ receptor distribution were generated. The reference used in the model was the mean time activity curve of the cerebellar grey matter (excluding the vermis), i.e. parts of the cerebellum that are devoid of specific 5-HT$_{1A}$ receptor binding (Parsey et al 2005).

1.2. MRAC methods

1.2.1. MaxProb method. Similarly to Mérida et al (2015) but with the new database, synthetic pseudo-CTs were generated for each subject in a leave-one-out design. Each subject’s MR image was used as a target, and the 39 remaining subjects as the atlas database. Pairwise nonrigid registration of each atlas MR image to the target MR image was computed and used for propagating the atlas CT into the target space. The pseudo-CT was generated through voxelwise atlas selection and intensity fusion. The pipeline of the MaxProb method is shown in figure 1; a detailed step-by-step description follows.

1.2.1.1. Registration. MR images from the original database of co-registered MR-CT pairs (see section 2.1.1) were mapped to each target subject’s MR image using affine registration, followed by non-rigid registration (NiftyReg suite: http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg) (Ourselin et al 2001, Modat et al 2010), based on cubic-B-spline, with normalized mutual information as similarity measure and a control point spacing of 5 mm. The transformations obtained from the MR non-rigid registration were then applied to the corresponding co-registered CT. This step yielded the registered database (figure 1, Step 1).
1.2.1. Atlas selection and fusion. Registered CT atlases were segmented in three tissue classes, defined by intensity thresholding on the CT images (Poynton et al 2014):

- Air: $<-500$ HU
- Soft tissue: $(-500; 300)$ HU
- Bone: $>300$ HU

For each voxel in the target subject space, majority voting was performed across the registered CT atlases to determine a majority tissue class label (figure 1, Step 2). If there was more than one modal value in the distribution, one of the equiprobable tissue classes was randomly selected (Hammers et al 2003). Finally, the voxel intensity value of the pseudo-CT was determined by averaging CT HU values of atlases belonging to the majority class for the corresponding voxel (figure 1, Step 2).

1.2.2. SingleAtlas method. As a point of reference, we developed a simplified method (SingleAtlas), which uses only one atlas (MR and CT pair), randomly selected from the database. The pseudo-CT is built by registering the atlas MRI to the subject space and then warping the CT image (similar to Schreibmann et al (2010)). The same registration parameters as for the multi-atlas approach were used. The transformed single CT constitutes the pseudo-CT. We compared the MaxProb to the SingleAtlas method to determine whether the complexity of the multi-atlas approach yielded any accuracy advantage.

1.2.3. Pseudo-CT background. The background of real CT images (i.e. atlas images in the context of leave-one-out evaluation) contained the pillow and other components that contributed to the attenuation in the PET images. To account for this additional attenuation, the background in the pseudo-CT image was replaced with the real CT image background for each subject. Note that this background issue will not need to be managed for PET-MR imaging with the mMR scanner, since the hardware $\mu$-map is integrated to the subject attenuation map by the manufacturer’s reconstruction software.

1.3. Evaluation

1.3.1. Pseudo-CT accuracy. Evaluation of the synthetized pseudo-CT was restricted to voxels within a head mask. Head masks were generated from the CT images as described in Merida et al (2015). Each generated pseudo-CT was compared to the subject’s real CT (ground truth CT). Real CT and pseudo-CT images were labelled by intensity thresholding (see thresholds above). The Jaccard overlap index (Jaccard 1901; intersection over union) was computed per tissue class (air, soft tissue and bone), and the percentage of misclassified voxels in the pseudo-CT compared to the real CT was employed as a metric reflecting the accuracy of the generated pseudo-CT. Various thresholds for tissue classification were tested in the evaluation and similar results were found (results not shown). In addition, we computed the mean absolute error (MAE) in Hounsfield units across the head mask (equation (1)).

$$\text{MAE} = \frac{\sum_{i=1}^{K} |pCT_i - rCT_i|}{K}$$

where $pCT_i$ refers to the value (in HU) for pseudo-CT at voxel $i$, $rCT_i$ refers to the value (in HU) for ground truth CT at voxel $i$ and $K$ is the total number of voxels in the volume of interest.

The SingleAtlas and MaxProb pseudo-CT generation methods were compared on these quality criteria using paired Wilcoxon signed-rank tests. The threshold of statistical significance
was set at a \( p \)-value of 0.05, divided by the number of comparisons (two in this study) to correct for multiple comparisons (Bonferroni 1936).

1.3.2. Impact on PET quantification. The error induced by the MRAC methods on PET quantification and binding parameters was assessed using the PET image reconstructed with real CT as the ground truth. The assessment was performed on activity values for PET \([^{18}\text{F}]\)FDG data, and on BP\(_{\text{ND}}\) images and time activity curves for \([^{18}\text{F}]\)MPPF data.

MR images and segmentation labels were registered to the corresponding PET images with SPM12 (Register function). \([^{18}\text{F}]\)FDG and BP\(_{\text{ND}}\) images were spatially normalized to MNI space by using SPM12.

The bias introduced by the MRAC methods was calculated as the relative error (equation (2)) or absolute error (equation (3)):

\[
\text{Relative error} (\%) = \frac{\text{PET}_{\text{MRAC}} - \text{PET}_{\text{CTAC}}}{\text{PET}_{\text{CTAC}}} \times 100 \tag{2}
\]

\[
\text{Absolute error} (\%) = \left| \frac{\text{PET}_{\text{MRAC}} - \text{PET}_{\text{CTAC}}}{\text{PET}_{\text{CTAC}}} \right| \times 100 \tag{3}
\]

where \(\text{PET}_{\text{CTAC}}\) refers to PET data corrected for attenuation with the ground truth CT and \(\text{PET}_{\text{MRAC}}\) is the PET data corrected with the MRAC \textit{SingleAtlas} or \textit{MaxProb} methods.

1.3.2.1. Regional analysis. Regional mean and standard deviations of \([^{18}\text{F}]\)FDG activity were extracted in the subject space for the brain tissue masks (GM, WM, CSF). Regional mean and standard deviations of \([^{18}\text{F}]\)FDG and \([^{18}\text{F}]\)MPPF BP\(_{\text{ND}}\) were also extracted in regions of the Hammers\textit{m}ith MRI segmentation (84 brain regions for \([^{18}\text{F}]\)FDG data and a selection of regions that exhibit specific \([^{18}\text{F}]\)MPPF binding, plus the cerebellum, masked by GM for \([^{18}\text{F}]\)MPPF data). The relative errors between the ground-truth PET and PET reconstructed with the MRAC methods were calculated for each ROI. Average bias for each cerebral region was computed per radioactive tracer, across the subjects. Statistical significance of the differences in regional evaluation between ground truth and MRAC methods was studied with a paired Wilcoxon signed-rank test.

1.3.2.2. Voxel-wise analysis. Voxel-wise parametric maps of the bias were computed for \([^{18}\text{F}]\)FDG PET data: for each subject, the image of relative error between \(\text{PET}_{\text{CTAC}}\) and \(\text{PET}_{\text{MRAC}}\) was calculated in the subject space. The images of relative error were then normalized to MNI space, averaged and finally masked with a brain mask.

Voxel-based analysis to assess differences between \(\text{PET}_{\text{CTAC}}\) and \(\text{PET}_{\text{MRAC}}\) for \([^{18}\text{F}]\)FDG PET was performed with SPM12 using an ANOVA with the factors methods and subjects. The resulting statistical parametric maps were thresholded at an uncorrected significance level of \(p < 0.001\) for illustration, and surviving clusters at a significance level of \(p < 0.05\) corrected for multiple comparisons (family wise error).

1.3.2.3. Kinetics. We investigated the effect of activity distribution on the accuracy of PET activity quantification with the MRAC methods. \([^{18}\text{F}]\)MPPF time activity curves (TAC) were extracted for several ROIs. The first frame was removed because of low count rates close to zero. The bias on TACs was computed by calculating the relative error frame by frame. The frame biases were then averaged across subjects.
2. Results

The computation time required to generate a pseudo-CT with the MaxProb method was around 1.5 h using a single core. Using a six-core machine, the multiple registrations required in the process can be parallelized, reducing the run-time to about 15 min.

2.1. Pseudo-CT results

In this section we report results obtained for the evaluation of the pseudo-CT generated with SingleAtlas and MaxProb MRAC methods (see section 2.2.1).

2.1.1. Qualitative results. Figure 2 shows the ground truth CT and pseudo-CTs generated with the SingleAtlas, and MaxProb methods for a randomly selected subject. The difference image between the real CT and each pseudo-CT (real CT — pseudo-CT) is also shown. Pseudo-CTs computed with both methods showed, in general, strong agreement with the ground truth CT. However, the error for the skull was much larger for the SingleAtlas method (error range from $-3000$ to $3000$ HU) than for the MaxProb method (error between $-500$ and $500$ HU). The small amount of air in the mastoid cells is not well reproduced in the MaxProb pseudo-CTs but misplaced in the SingleAtlas approach.

2.1.2. Quantitative evaluation. Volumes of the various head tissues and non-tissue components with dissimilar attenuation properties, i.e. air, soft tissue, and bone (‘tissues’, mean ± standard deviation) within the head mask for all real CTs of the database were $180 \pm 36 \text{cm}^3$ for air, $2309 \pm 267 \text{cm}^3$ for soft tissue and $626 \pm 95 \text{cm}^3$ for bone. A box plot of the Jaccard indices computed per tissue class and methods is shown in figure 3. Mean values, standard deviations, and results from the statistical comparisons are summarized in table 1. The mean
Figure 3. Boxplots of Jaccard index per tissue class and per method. Note that y axis scales differ between plots. Centre lines correspond to medians, boxes to interquartile ranges, and whiskers to robust ranges. Outliers are represented as dots. Note that Jaccard indices obtained with the SingleAtlas approach were systematically lower than with the MaxProb method.

Table 1. Jaccard index (mean ± standard deviation) per method and per tissue class. Paired Wilcoxon signed-rank test (*p < 0.05 MaxProb versus SingleAtlas).

<table>
<thead>
<tr>
<th>Tissue</th>
<th>SingleAtlas</th>
<th>MaxProb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>Air</td>
<td>46.36 ± 4.96</td>
<td></td>
</tr>
<tr>
<td>Soft tissue</td>
<td>84.91 ± 1.43</td>
<td></td>
</tr>
<tr>
<td>Bone</td>
<td>66.78 ± 3.32</td>
<td></td>
</tr>
</tbody>
</table>

Paired Wilcoxon signed-ranked test (*p < 0.05)

Jaccard index obtained with the SingleAtlas method was 46.36% for air, 66.78% for bone and 84.91% for soft tissue. The MaxProb method systematically performed better, with a Jaccard index of around 10 points above the SingleAtlas method for air and bone. Paired Wilcoxon signed-rank tests showed that all the differences were statistically significant (table 1).

The percentage of voxel classification error (mean ± standard deviation) across all subjects, per method and error type, is reported in table 2. ‘Bone_as_air’ means that a voxel was classified as air in the pseudo-CT when it should have been bone according to the ground truth CT; the remaining row labels are formed in the same manner. Errors are expressed as the percentage of the voxels within the head mask. The total classification error was approximately 12.3% for the SingleAtlas method and decreased to around 8.5% for the MaxProb method. Significantly better performance was achieved with the MaxProb method compared to the SingleAtlas approach.

Mean absolute errors of pseudo-CT intensities computed voxel-by-voxel, on the head mask per tissue class and for the global head volume, across all subjects are shown in table 3. Significantly smaller errors were obtained for the MaxProb method compared to the SingleAtlas approach.
2.2. Results for static PET data

In this section we present the results obtained with the MRAC methods tested in terms of accuracy of the resulting reconstructed image. Differences between OP-OSEM3D and FBP2D were negligible. Therefore only results obtained with the iterative reconstruction algorithm are shown in their entirety. Parts of the results obtained with the filtered back-projection algorithm are provided in supplementary material for comparison.

2.2.1. Global evaluation. Figure 4 shows the absolute error (in %) between PET CTAC and each PETMRAC for GM, WM and CSF for static [18F]FDG PET. Table 4 reports both relative and absolute biases per method and tissue class. For a given tissue class, both methods had similar average results. The mean and standard deviation of absolute bias were slightly, but not significantly higher for the SingleAtlas method than for MaxProb. Results also revealed that the performance discrimination between SingleAtlas and MaxProb is less obvious using metrics computed from the reconstructed images than when they are directly computed from the generated pseudo-CT (difference to the MaxProb method around 10 points using the Jaccard index on the pseudo-CTs versus less than one point using the reconstructed PET images).

2.2.2. Regional evaluation. The mean regional bias (in %) measured for the 84 brain structures from [18F]FDG images reconstructed with the SingleAtlas and MaxProb MRAC methods is shown in figure 5.

With [18F]FDG data, the mean bias obtained with the MaxProb method was mostly between 0 and 1% (range from −0.6 to 2.5%), with a slight tendency for overestimation rather than
underestimation. Few structures had a bias larger than 2%: lateral part of anterior temporal lobe, middle and inferior temporal gyrus, fusiform gyrus and anterior orbital gyrus. Errors were larger with SingleAtlas (range from −2 to 4%) with a high variability that included over- and underestimations. There was substantial localized bias in all regions of the parietal lobe.

2.2.3. Local evaluation. Figure 6 shows the mean voxel-wise difference images (averaged across the 23 subjects in the standard space) between real CT and pseudo-CT (pseudo-CT—real CT) as well as the mean voxel-wise relative error between the [18F]FDG images reconstructed with the real CT and those obtained with SingleAtlas and MaxProb.

The largest errors in synthetic pseudo-CTs can be seen at the air and bone boundaries. We note that due to the nature of the reconstruction process, there is no direct translation between inaccuracies in the CT image and the resulting bias in the reconstructed image. However, in general, underestimation of the pseudo-CT value leads to underestimation of the activity in
the surrounding brain tissues in the PET image, and overestimation of the pseudo-CT value leads to overestimation of the activity in the surrounding brain tissues in the PET image. The SingleAtlas approach produces errors up to ±1000 HU in mean image difference. These inaccuracies were localized in bone and air-filled regions. [18F]FDG PET data reconstructed with the SingleAtlas pseudo-CT showed local bias around -2 and 2% inside the brain. In some brain regions, the bias was greater than 25%, for example in the postcentral gyrus, the superior frontal gyrus and the cerebellum. With the MaxProb method, the error at the voxel level was near 0% throughout most of the brain. Error was predominantly localized in the regions of the brain adjacent to the skull, whereas bias was less substantial in the centre of the brain. Some
small clusters of voxels approached 15% error, namely in the inferior temporal lobe, anterior orbital gyrus and the cerebellum. However the cerebellar errors become negligible when averaged over the cerebellar region of interest as a whole (see figures 5 and 7).

2.2.4. SPM analysis. The t score maps of the difference between PET_CTAC and the different PET-MRAC for the [18F]FDG, thresholded at $p < 0.001$, are shown in figure 7. [18F]FDG activity image corrected with the SingleAtlas AC showed more extensive regions of overestimation with higher...
t-scores (figures 1(a) and 7) than those seen with MaxProb AC (figures 2(a) and 7). Regions affected were the main part of parietal lobe, brain stem and orbitofrontal cortex. \[^{18F}\]FDG images corrected with the SingleAtlas AC were also affected by underestimation in disparate regions such as the frontal lobe and the cerebellum (figures 1(b) and 7). Data corrected with MaxProb showed overestimation in few small clusters that were localized in the frontal lobe and orbitofrontal regions (figures 2(a) and 7). No significant regions of underestimation were found (figure 7(a) and 7).

The illustrative statistical parametric maps shown in figure 7 were then thresholded at a significance level of $p < 0.05$ corrected for multiple comparisons, and only clusters with an extension exceeding the expected size ($\langle k \rangle = 5.75$) were retained. SingleAtlas yielded 43 significant clusters, covering large areas of the brain; slightly more than half were overestimations. In the cerebellum, the reference region for a number of PET tracers, eight significant clusters of overestimation were found with t-scores up to 7.2, and four significant clusters of underestimation with t-scores up to 6.8. The significant clusters represented more than 328 cm$^3$ (20%) of the analysed brain volume (1655 cm$^3$). In contrast, only two significant clusters of overestimation were found with MaxProb MRAC (in the left anterior orbital gyrus, $z = 5.3$, 223 mm$^3$, and right superior frontal gyrus, $z = 5.2$, 96 mm$^3$), and no clusters of underestimation. This represented 0.02% of the analysed brain volume.

2.2.5. Outliers. We found two significant outlier subjects in the \[^{18F}\]FDG PET analysis, who had a regional bias exceeding 10% for some labels when MaxProb MRAC was used (five outliers...
with the SingleAtlas approach). Figure 8 shows PET data corrected with the MaxProb pseudo-CT. For the first subject, an abnormally strong signal in the PET image was observed around the anterior frontal cortex bilaterally, with a gradient in PET bias from the bone towards the centre of the brain. The other subject had a substantial overestimation in the bias map localized around the lateral part of the left temporal lobe, but no signal increase was visible on the PET image.

2.3. Results for dynamic PET data

This section describes the impact of the attenuation correction approach on the quantification of dynamic $^{18}$F]MPPF PET data and the subsequent kinetic modelling. As in the previous section, OP-OSEM3D and FBP2D reconstruction algorithms showed similar tendencies in the dynamic evaluation, therefore only results obtained with the iterative reconstruction
algorithm are shown in their entirety. Part of the results obtained with the filtered back-projection algorithm are provided in supplementary material for comparison.

2.3.1 Regional evaluation. Results for $^{18}$F]MPPF are shown in figure 9 for a selection of 44 regions that exhibit specific $^{18}$F]MPPF binding (Costes et al 2002). To investigate the effects of the inaccuracies produced by the tested MRAC methods on the kinetic modelling step, we present the mean bias computed from the late static $^{18}$F]MPPF images (generated by time-weighted averaging over the final ten minutes of the acquisition; shown as dotted lines), as well as the mean bias for the BPND images (as solid lines).

The bias measured from late static $^{18}$F]MPPF images was higher than for $^{18}$F]FDG, with the mean error ranging from $-1.2$ to $5.0\%$ for MaxProb. As for the $^{18}$F]FDG evaluation, a general overestimation for most of the assessed regions was also noted. SingleAtlas showed bidirectional, but mainly negative, bias (from $-8.5$ to $1.7\%$) for the tested brain structures. The variability of the bias was again higher than for the MaxProb method.

Bias on the BPND ranged from $-2.5$ to $5.0\%$ for MaxProb. For SingleAtlas, the maximum negative bias reached $-9.3\%$ and the positive bias reached $3.3\%$. Two regions were particularly strongly affected by errors: lateral orbital gyrus ($-9.3\%$) and lateral part of anterior temporal lobe ($-7.3\%$). One can note that BPND images, error patterns did not correlate well with errors in activity values measured in the late static $^{18}$F]MPPF images.

Figure 10 shows an example of $^{18}$F]MPPF PET image at early and late-phase of the dynamic acquisition for a randomly selected subject. An example of a late-phase $^{18}$F]FDG PET image (from a different subject) is shown for comparison. Note the very different activity distribution in the brain, depending on the tracer used and the time of acquisition relative to injection.

2.3.2. Evaluation on time activity curves. The analysis of the measured TACs showed that the mean bias computed from the early static PET image (corresponding to the first three minutes of the dynamic image, during the perfusion of tracer in the tissues) across the 44 investigated regions was $0.2 \pm 0.8\%$ for SingleAtlas and $1.0 \pm 0.7\%$ for MaxProb, while when computed from the late static PET image (the last ten minutes of the dynamic image) and across the same regions the mean bias was $-2.6 \pm 2.2\%$ for SingleAtlas and $1.5 \pm 1.1\%$ for MaxProb, suggesting first that inaccuracies in AC maps impacted the reconstructed time frames differently and that, overall, the resulting error seemed to increase in magnitude over time with the two MRAC methods tested.

Mean errors computed as a function of time for selected regions that were obtained with SingleAtlas and MaxProb are shown in figure 11. The graph reports bias on the time-activity
Figure 11. Mean bias over time for the PET frames across subjects for SingleAtlas and MaxProb MRAC methods, i.e. during the 60 min $[^{18}F]$MPPF PET acquisition, for selected representative regions (see text). Resulting BP$_{ND}$ biases are also given (left panel). There is no evident relation between the bias on TACs and the bias on regional BP$_{ND}$.
curve across time for the cerebellum excluding vermis (no specific binding, reference region for modelling), hippocampus (high binding of [18F]MPPF), anterior temporal lobe lateral part (a region where both MRAC method had substantial bias on BPND) and posterior temporal lobe (a region where both MRAC methods had weak bias on BPND). Only labels in left hemisphere are shown here but similar curves were obtained for the labels in the right hemisphere. Overall, the magnitude of the mean bias tended to increase during the measurement period. This was the case for most of the 44 studied regions. However, the magnitude of bias was lower and less time dependent for the MaxProb approach than for the SingleAtlas method.

For the cerebellum, the reference region used for [18F]MPPF modelling, the bias on TACs fluctuated over time between $-2$ and $2\%$ for SingleAtlas and 0 and $2\%$ for MaxProb. The bias tended to increase slightly over time, in particular for SingleAtlas. In the hippocampus, both methods yielded very low and almost constant bias over time and negligible bias for BPND. For the lateral part of anterior temporal lobe, the magnitude of bias increased over time, with a higher slope in the case of SingleAtlas. In this region, the bias reached 4% during the last ten minutes of the acquisition for the MaxProb method and $-7.5\%$ for SingleAtlas. The resulting BPND were also strongly biased for both MRAC methods. A large increase in bias was observed in the posterior temporal lobe, in particular for the SingleAtlas method. However, the biases obtained for BPND were very close to 0% for both MRAC methods.

Figure 12 shows, for each of the seven subjects, the mean bias (in %) averaged across the 44 brain regions, per subject and per MRAC method as a function of the coefficient of variation of tracer activity. Time-dependency is shown in levels of grey. Linear regression was performed per subject. The shade around linear regression shows the standard error of the slope (confidence interval of 95%).
yielded higher COV and bias. MaxProb yielded time activity measurements with errors that were significantly smaller dependent on the activity distribution. This was confirmed by the computation of the mean slope across the seven subjects included in the study which was $-7.0$ for SingleAtlas and $1.2$ for MaxProb (figure 12).

The plot of the slopes obtained by linear regression of the data presented in figure 12, versus the global mean absolute error (MAE) between the ground truth CT and pseudo-CT computed within the head mask for SingleAtlas and MaxProb methods is shown in figure 13. The figure suggests a correlation between the dependence of the bias with the image uniformity, expressed with the slope, and the quality of the generated pseudo-CT. It also shows that MaxProb generated more accurate pseudo-CT than SingleAtlas, yielding dynamic activity measurements with a bias that was less dependent on the activity distribution.

3. Discussion

3.1. Main results

Pseudo-CTs could be generated with acceptable accuracy with MaxProb. SingleAtlas produced larger errors as well as many more outliers than the multi-atlas approach, a finding that we expected because a single image cannot reflect inter-subject variability (Heckemann et al 2006).

In this paper, we used a larger atlas database ($n = 40$ MRI-CT pairs) than in our previous work (Mérida et al 2015, $n = 27$ pairs), with more thinly spaced CT slices. The results were very similar. This is important, as it suggests that the multi-atlas method performs similarly across databases.

We showed that both regional and local evaluation of the errors is relevant, since large bias localized in regions near bone may become averaged and not detected in a global and even a regional evaluation. This point is important for studies in neuroscience that focus on small brain structures, but also for clinical studies, e.g. in the presurgical evaluation for epilepsy where small cortical abnormalities are sought. The results obtained with the different metrics showed strong coherence.
3.2. Remaining errors on pseudo-CT images

In the MRAC methods, attenuation of brain structures that have boundaries with air cavities or with bone is difficult to estimate because of the low contrast between air and bone in the MR images. Figure 6 shows that CT values in such regions are particularly biased for the SingleAtlas method, but are well managed in the multi-atlas approach despite a slight local overestimation of around 300 HU. The effect of error on bone was generalized in the SingleAtlas AC method, and these inaccuracies produced both overestimations and underestimations on PET images all around the cortex. The small residual bias observed on the parametric image for MaxProb (figure 6) may be partially affected by a lack of information around the neck in the atlas database. A new database with extended field of view is under investigation.

3.3. Analysis of dynamic data

PET-MR scanners are currently largely used for research, and full quantification with kinetic modelling will often be required.

To our knowledge, this work highlights for the first time that inaccurate attenuation maps introduce bias in measured TACs that depends on the spatial distribution of the tracer in the head. Note that a similar finding has recently (and independently to this work) been reported in the context of PET/CT lung imaging (Holman et al 2016). In dynamic acquisitions, the spatial distribution of the tracer within the brain can change across time from being rather uniform (early frames: blood flow/perfusion) to very contrasted (late frames: specific binding) (see figure 10). In this situation, inaccurate attenuation maps will not only bias the measured TACs in magnitude but also in shape with unpredictable repercussions at the kinetic parameter computation step. In addition, with inaccurate attenuation maps varying performance is to be expected for late static images for tracers with heterogeneous distribution in the brain, which we have shown with late [18F]MPPF compared to late [18F]FDG uptake. Results presented in figures 12 and 13 illustrate this finding. The reconstruction step is a complex process. But if the activity value present in a single voxel of a reconstructed image is conceptualized simply as a linear combination of projection bin values corrected (e.g. multiplied) by their associated attenuation correction factors, it is easy to perceive the link between count distribution in projection space and bias in voxel values when attenuation correction factors are inaccurate. If the count distribution was uniform, the voxel value would not be very biased even given slightly incorrect attenuation correction factors. A mathematical framework that describes the quantification error in the PET image due to an inaccurate $\mu$-map has been introduced in Thielemans et al (2008). In our study, the results obtained with the dynamic evaluation on [18F]MPPF PET data showed a strong similar tendency (figure 12) for six of seven subjects assessed.

Note that other phenomena could explain the dependence of PET bias on tracer spatial distribution: changing counting statistics across time can influence the reconstruction algorithm as well as the scatter correction. In this work we showed that this dependence was purely the result of an inaccurate $\mu$-map, by verifying the observations reported in figure 11 with the same data reconstructed with FBP2D. This confirmed that the biases did not depend on the reconstruction method, and in particular that their evolution across time was not due to convergence properties that can vary with changing counting statistics when using iterative reconstruction methods.

We also verified that this evolution was not caused by inaccuracies introduced during the scatter correction, a step that uses the attenuation map and whose performance could be
influenced by the statistics and activity distribution within each emission time frame. The
dynamic PET data from a single subject were reconstructed without scatter correction using
the ground-truth CT and SingleAtlas pseudo-CT as the AC method. Results (see supplementary
material) showed that the scatter correction only explains a small part of the error and its
evolution across time, a finding which is supported by results reported by Burgos et al (2014b)
in the context of static PET acquisitions.

The MAE values obtained for pseudo-CT evaluation (Figure 13) were consistent with
those reported in Burgos et al (2014). MaxProb values were equivalent to those of the multi-
 atlas method described in Burgos et al (2014), and SingleAtlas had similar MAE scores as
the AC method based on the UTE image (the vendor’s method implemented on the scanner).
This suggests that the UTE method can also lead to considerable bias across time in dynamic
data.

The cerebellum contains few receptors belonging to the most frequently studied neuro-
transmitter systems, and it is relatively spared in neurodegenerative disease. Therefore, it is
routinely used as a reference region for modelling or internal standardization (e.g. stand-
ard uptake value ratios). Correct quantification of cerebellar radioactivity concentrations is
therefore particularly important but had not been obtained with standard vendor implementa-
tions (Andersen et al 2014). We argue that this problem has now been solved with multi-
atlas approaches. The local errors observed in the cerebellar region with the MaxProb method
(figure 6) averaged out over the entire cerebellum (figure 5) and did not affect the kinetic
modelling (see figures 9 and 11). \( BP_{ND} \) computation relies on the complex modelling of the
activity concentration over time. We found no obvious correlations between biases in activity
estimates and biases in the resulting \( BP_{ND} \).

3.4. Outliers

We found an anatomical explanation for the unusually large errors in PET data seen in
two subjects locally: outlier #1 had abnormally large frontal sinuses and outlier #2 had
undergone a lobectomy, and a craniotomy in the lateral skull overlying the temporal lobe.
These characteristics can be observed on the real CT images. The multi-atlas methods did not
handle these anatomical abnormalities well. However, bias remained localized to the immedi-
ate vicinity of the pseudo-CT abnormalities (figure 8). The cluster showing high quantification
error for outlier subject #2 (figure 8) was located in the CSF, explaining why no difference
is visually discernible on the PET images (figure 8). The outlier detection was based on the
regional quantification, which did not handle the postoperative change correctly, while there
was no relevant error propagation into the brain image itself.

The signal increase resulting from anatomical variants or postoperative changes may some-
times be clinically pertinent and would present a risk for misdiagnosis if the AC error was not
noticed. In our case, the errors in the pseudo-CTs can be predicted from visual review of the
MR image (large frontal sinuses can be seen on T1-weighted MRI) for outlier #1 and from
the medical history (craniotomy) for outlier #2. In PET-MR imaging, similarly to PET/CT but
possibly even more so, referring clinicians should take care to provide an accurate request that
lists all pertinent history. Similarly to PET/CT, reporting clinicians have to be aware of possible
pitfalls, and should be able to simultaneously visualize the PET image, the MR image,
and the attenuation map, and perhaps explicitly check for unusually large sinuses or fluid-filled
sinuses (for which we did not have examples in our cohort). It is to be expected that reporting
clinicians will become accustomed to PET-MR artefacts, just as they already are accustomed to
interpreting e.g. FLAIR artefacts, or CT artefacts due to bone or metal. For now, areas close to
craniotomies may be difficult to interpret, and clinicians should remain aware of the possibility
of large sinuses causing local artifacts. In some special cases, e.g. in the case of brain tumours with craniotomies, it may be preferable to plan for an additional low-dose CT scan.

The MaxProb method might be further improved by deriving subject-specific bone information from the MR images, for example via UTE sequences (Roy et al 2014, Ladefoged et al 2015), and combining it with the multi-atlas pipeline. It may also be possible to explore the number of atlases selected during the MaxProb fusion process and use this to determine optimal numbers of (similar) atlas pairs in the database and/or selection of atlas pairs. Another area for exploration is the effect of the degree of smoothing of the pseudo-CT—which by construction will be smoother than the real CT—in the reconstruction process and the impact on quantitative PET analysis. The method could also in principle be refined by simply updating the coordinate system rather than reslicing the MR in CT space.

No UTE data was available in this study. Prior work has shown superiority of multi-atlas methods over UTE for AC (Burgos et al 2014a). Nevertheless, it is likely that further development will take place around specialized MR sequences for radiodensity mapping and that combining a tailored sequence with an atlas-based algorithm may be optimal.

4. Conclusion

Our findings entail the following points that we consider consequential. 1) Knowledge about the correlation between MR signal intensity and tissue attenuation should be drawn from a multi-subject reference database – using single-subject reference correlation data leads to inferior attenuation maps. 2) Inaccuracies in attenuation map estimates lead to biases that depend on the spatial distribution of the activity, which can be problematic in dynamic PET imaging involving kinetic modelling. 3) Multi-atlas attenuation correction provides highly accurate data that are largely equivalent to data corrected via CT. The method presented here entails little bias in static PET activity estimation or in the BPND maps generated through compartmental modelling and should therefore be of sufficient quality and robustness for research applications. It is also expected that PET-MR images corrected via multi-atlas AC can be used clinically, provided subject characteristics like prior surgery or anatomical variants are duly considered.

Software to implement our method is available for research purposes.

Acknowledgments

This work was supported by the following: LILI-EQUIPEX—Lyon Integrated Life Imaging: hybrid PET-MR ANR project-11-EQPX-0026, the French National Agency for Research and Technology (ANRT), CESAME - Brain and Mental Health ANR-10-IBHU-0003 program, Neurodis Foundation, and Siemens Healthcare SAS France. We would like to thank Siemens Medical Solutions, Knoxville, USA for providing the PET reconstruction software e7tools. We thank our colleagues at the CERMEP Lyon for help in the acquisition and transferring of PET data; and Marjorie Villien for valuable advice on this work.

References

Andersen F L et al 2014 Combined PET/MR imaging in neurology: MR-based attenuation correction implies a strong spatial bias when ignoring bone Neuroimage 84 206–16


2857
Mehranian A, Arabi H and Zaidi H 2016 Quantitative analysis of MRI-guided attenuation correction techniques in time-of-flight brain PET/MRI Neuroimage 130 123–33