ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 32 (2006) 1591 – 1607

Comparative assessment of statistical brain MR image segmentation algorithms and their impact on partial volume correction in PET

Habib Zaidi,* Torsten Ruest, Frederic Schoenahl, and Marie-Louise Montandon

Division of Nuclear Medicine, Geneva University Hospital, CH-1211 Geneva 4, Switzerland

Received 27 October 2005; revised 28 April 2006; accepted 10 May 2006
Available online 7 July 2006

Magnetic resonance imaging (MRI)-guided partial volume effect correction (PVC) in brain positron emission tomography (PET) is now a well-established approach to compensate the large bias in the estimate of regional radioactivity concentration, especially for small structures. The accuracy of the algorithms developed so far is, however, largely dependent on the performance of segmentation methods partitioning MRI brain data into its main classes, namely gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). A comparative evaluation of three brain MRI segmentation algorithms using simulated and clinical brain MR data was performed, and subsequently their impact on PVC in ^{18}F-FDG and ^{18}F-DOPA brain PET imaging was assessed. Two algorithms, the first is bundled in the Statistical Parametric Mapping (SPM2) package while the other is the Expectation Maximization Segmentation (EMS) algorithm, incorporate *a priori* probability images derived from MR images of a large number of subjects. The third, here referred to as the HBSA algorithm, is a histogram-based segmentation algorithm incorporating an Expectation Maximization approach to model a four-Gaussian mixture for both global and local histograms. Simulated under different combinations of noise and intensity non-uniformity, MR brain phantoms with known true volumes for the different brain classes were generated. The algorithms' performance was checked by calculating the kappa index assessing similarities with the "ground truth" as well as multiclass type I and type II errors including misclassification rates. The impact of image segmentation algorithms on PVC was then quantified using clinical data. The segmented tissues of patients' brain MRI were given as input to the region of interest (RoI)-based geometric transfer matrix (GTM) PVC algorithm, and quantitative comparisons were made. The results of digital MRI phantom studies suggest that the use of HBSA produces the best performance for WM classification. For GM classification, it is suggested to use the EMS. Segmentation performed on clinical MRI data show quite substantial differences, especially when lesions are present. For the particular case of PVC, SPM2 and EMS algorithms show very similar results and may be used interchangeably. The use of HBSA is not recommended for PVC. The partial volume corrected activities in some regions of the brain show quite large relative differences when performing paired analysis on 2 algorithms, implying a careful choice of the segmentation algorithm for GTM-based PVC.

Keywords: MRI; PET; Brain imaging; Segmentation; Partial volume effect

Introduction

During the last decade, neuroimaging has advanced elegantly in the medical and research arenas. Molecular brain imaging using positron emission tomography (PET) plays a valuable role in the assessment of cellular targets, thus providing clinicians and neuroscientists with relevant information in various pathologies and neurological disorders (Zaidi and Montandon, 2006). Nevertheless, PET is obviously not the only major non-invasive tool for the assessment of brain disease. Major new technologies, such as spiral computed tomography (CT), high-field magnetic resonance imaging (MRI), bioluminescent and fluorescent imaging, and many other technologies, have now blurred the artificial distinction that once set PET apart as a "functional" rather than "anatomic" imaging modality. Nonetheless, PET maintains an exclusive standing in the delivery of targeted therapies, but its superior picomolar sensitivity is being challenged by competing technologies.

The high contrast of MRI makes the method of choice to detect abnormalities in the brain in addition to offering the possibility of partitioning the brain into its main classes. Automated medical image segmentation is becoming an increasingly important image processing step for a number of clinical and research applications including but not limited to brain volumetry, treatment planning in radiation therapy, surgical planning, and image-guided intervention procedures. Moreover, modern molecular brain imaging using PET relies on high-resolution segmented anatomical data for anatomically guided statistical image reconstruction (Baete et al., 2004), attenuation compensation (Zaidi et al., 2003), and partial volume effect correction (PVC) (Rousset and Zaidi, 2005; Rousset et al., 1998). The performance of those algorithms depends largely on the quality of the segmentation output, and thus special attention

* Corresponding author. Fax: +41 22 372 7169.
E-mail address: habib.zaidi@hcuge.ch (H. Zaidi).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2006.05.031

should be given to the segmentation procedure and its algorithmic implementation.

Image segmentation has been identified as the key problem of medical image analysis and remains a popular and challenging area of research. A wide variety of brain MR image segmentation techniques including a number of promising approaches were devised and are described in the literature (Clarke et al., 1995; Suri et al., 2002; Thacker et al., 2004). This includes thresholding, region growing, classifiers, clustering, edge detection, Markov random field models, artificial neural networks, deformable models, atlas-guided, and many other approaches (Boudraa and Zaidi, 2005). The algorithms devised specifically for segmenting the cortex from 3D MR images fall within two broad categories: voxel classification and deformable models (Duncan et al., 2004). Automated segmentation approaches have proven sufficiently accurate for volumetric display and may also be adequate for brain volumetry. The inclusion of bias field correction and partial volume effects in the segmentation paradigm has especially proven valuable (Viergever et al., 2001). Nevertheless, more comprehensive studies of brain anatomy and physiology will necessitate more sophisticated segmentation approaches. More recent atlas-based approaches include *a priori* models of lesion growth to allow accurate brain tissue segmentation even in pathological brains and when space-occupying lesions are present (Cuadra et al., 2004; Pollo et al., 2005). Some investigators performed comparative assessment studies of MRI segmentation approaches focusing mainly on cortex segmentation and tissue volume computation using simulated and real data (Grau et al., 2004; Cuadra et al., 2005).

Early attempts to compensate for PVE date back to the time where they were first pointed out as a serious limitation in quantitative analysis (Hoffman et al., 1979). Some correction methods require only the PET emission data. This includes pioneering research attempting some sort of PVC through the application of recovery coefficients described in the reference above as well as those derived from Kessler's formulation (Kessler et al., 1984). A distinct class of correction methods requires the definition of the various objects being imaged in addition to the characterization of the scanner's point spread function (PSF). These include anatomy-based post-reconstruction correction methods that make use of concomitant high-resolution structural information from MRI or CT (Rousset and Zaidi, 2005). It is worth emphasizing that the geometric transfer matrix-based method (Rousset et al., 1998) belonging to this class of techniques and used in this work is among the most popular techniques in the field.

For the particular application of MRI-guided PVC in brain PET, the main segmentation task is the extraction of cerebral tissue classes, namely gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). The accuracy of the PVC algorithm is highly dependent on the degree of accuracy achieved by the MRI-PET realignment and the MRI segmentation procedures. Several studies reported the impact of segmentation errors on the accuracy of PVC in both phantom and simulation studies (Meltzer et al., 1999; Strul and Bendriem, 1999; Frouin et al., 2002; Quarantelli et al., 2004; Rousset and Zaidi, 2005), however, there is a lack of detailed investigations of the effect of the MRI segmentation algorithm on PVC using clinical data where the ground truth is unknown. The objectives of this work are two-fold: firstly to assess qualitatively and quantitatively the performance of three commonly used brain MR image segmentation algorithms using simulated phantoms under controlled noise and intensity non-uniformity conditions and clinical data acquired in realistic conditions; and secondly to assess the impact of segmentation algorithms on a region-based PVC technique using clinical ^{18}F-FDG and ^{18}F-DOPA PET studies by using the output of each segmentation technique as input to the PVC algorithm. The three segmentation methods assessed in this work include the segmentation algorithm bundled in the Statistical Parametric Mapping (SPM2) package (Ashburner and Friston, 1997), the Expectation Maximization Segmentation (EMS) algorithm (Van Leemput et al., 1999b), which incorporate *a priori* probability images derived from MR images of a large number of subjects, and the classical approach for brain extraction and automatic tissue segmentation of MR images using the Expectation Maximization (EM) algorithm to model a four-Gaussian mixture for both global and local histograms (Kovacevic et al., 2002) here referred to as histogram-based segmentation algorithm (HBSA).

Materials and methods

Image segmentation algorithms

Given the wide range of MRI brain segmentation methods, the two main criteria instigating the choice made by the authors are software reliability and availability to the neuroimaging community and its applicability to MRI-guided PVC in PET. SPM is now considered among the gold standard tools and is being used world-wide both in research and clinical neuroimaging investigations. EMS is an open source code which can be freely downloaded from the web site of its authors (following registration), whereas HBSA can be obtained from its authors. The three techniques consider the three tissue classes of interest for PVC in PET, namely GM, WM, and CSF. Although the algorithms consider several clusters including scalp, eyes, background, etc., the segmentation output is limited to three classes. SPM2 was used as toolbox, whereas the two other packages required slight modifications beforehand to match our needs and image formats. Among the three algorithms assessed, only HBSA takes into account partial volume mixture voxels whereas EMS has the particularity of considering local spatial information by means of a Markov Random Field.

SPM2

The approach used in the segmentation algorithm incorporated in SPM2 (Ashburner and Friston, 1997, 2000) has some similarities with the EMS technique in the sense that both of them are based on the statistical brain atlas incorporated in SPM2 and use a two-step EM algorithm (see below). The affine transformation implemented in SPM2 is used to map the images and templates of the same modality. Basically, a maximum-likelihood (ML) clustering algorithm is used, which takes advantage of *a priori* images containing the statistical probability for each voxel to belong to GM, WM, or CSF. This information is used to initialize the algorithm and is also involved in computing the posterior probability. The algorithm then repeats a set of steps to compute the probabilities and estimate the cluster parameters. The parameters for the Gaussian mixture model, i.e. the number of voxels belonging to each cluster, the mean intensity of voxels within their respective cluster, and the variance in intensity of each cluster are estimated. Based on normally distributed voxels in the GM, WM, and CSF classes, respectively, and their currently estimated parameters, it is possible to assign to each voxel in each class a probability to belong to this class. The

prior knowledge of the spatial distribution from the probabilistic atlas for each tissue class is taken into account. Multiplying all the probabilities provides the likelihood, which the algorithm tries to maximize. Convergence is reached when the difference in likelihood between two subsequent iterations becomes negligible. In fact, the algorithm uses an EM approach (see below) where the E-step is dedicated to the computation of the probabilities and the M-step estimates the cluster parameters. Unlike previously developed partial volume mixture modeling approaches, the model adopted recently as an improvement of the method used in this work (SPM2) simply assumes that the intensity distribution of each class may not be Gaussian and assigns belonging probabilities according to non-Gaussian distributions (Ashburner and Friston, 2005).

EMS

The expectation–maximization segmentation (EMS) algorithm (Van Leemput et al., 1999a,b) attempts to improve iteratively the parameters of a mixture of normal distributions model corresponding to WM, GM, and CSF by interleaving an expectation and a maximization steps. In the expectation step, each of the voxels is classified into one of the tissue classes according to the current estimates of the normal distribution mixture model whereas in the maximization step, the normal distributions parameters are re-estimated, according to the current classification. This procedure ensures the increase of likelihood after each iteration. The algorithm is initialized by a statistical brain atlas incorporated in SPM2 (Ashburner and Friston, 1997, 2000) containing information about the *a priori* expected location of tissue classes, thus allowing full automation. In addition, it provides spatial information about the localization of one given tissue class, further constraining the EM algorithm. The image registration procedure that maps between the templates and images is performed using the method based on maximization of mutual information (Maes et al., 1997). An important difference between this and the other two segmentation algorithms is that it is the only one incorporating bias field estimation in the intensity model of brain tissue classes and its correction using the algorithm developed by the same authors (Van Leemput et al., 1999a).

HBSA

The segmentation algorithm here referred to as the histogram-based segmentation algorithm (HBSA) uses multispectral MRI sequences where the proton density and T2-weighted images are used to generate an intracranial brain mask for extraction of brain tissues and the T1-weighted image is used to segment brain tissues (Kovacevic et al., 2002). In contrast to EMS and SPM2 segmentation algorithms, this method is independent of *a priori* knowledge with respect to the tissues to be partitioned from the MR image. As a consequence, there is no need to coregister the MR image to be segmented to a stereotactic space. The segmentation algorithm consists of two steps. In the first step, it fits a four-Gaussian mixture model to a previously normalized global histogram of the T1-weighted MR image using the EM algorithm. The initialization of the EM is based on initial mean values of the GM, WM and CSF and a partial volume mixture of GM/CSF where automation is achieved by identifying the consistencies among images varying in intensity ranges as described in Kovacevic et al. (2002). The second step applies the parameters from the first step to initialize the fitting of the so-called local histograms. To obtain the local histograms, the entire MR image to be segmented is tiled into anisotropic volumes, each containing a small central core box. The anisotropic volumes are used to estimate the model parameters (weights, means, and standard deviations of the 4-Gaussian). The intensity cut-offs are calculated from the mean values obtained from the global histogram fitting and are defined as the average of CSF and GM (C_{cg}) and WM and GM (C_{wg}), respectively. Segmentation is performed exclusively on the central core box, where a voxel with a given intensity x is segmented as CSF, if $0 < x < C_{cg}$, as GM if $C_{cg} \leq x < C_{wg}$, and as WM if $C_{wg} \leq x$.

Simulated MRI brain phantoms

Since the noise and intensity non-uniformity (INU) cannot be controlled in clinical MR studies, and likewise the true volume and spatial localization of the three main brain classes are unknown, simulated MR images where the "ground truth" is known were used for assessment of the three segmentation algorithms described above. It should be noticed that unknown proportions of noise or INU in an MR image make it difficult to evaluate precisely the behavior in such cases. The digital brain MRI simulator (Collins et al., 1998; Kwan et al., 1999) developed at Montreal Neurological Institute (MNI, Canada) is widely used for the evaluation of segmentation algorithms (Ashburner and Friston, 2000; Grabowski et al., 2000; Kovacevic et al., 2002; Lemieux et al., 2003). Different pulse sequences may be chosen allowing multispectral MR data to be generated. Likewise, different options to contaminate the data are available, that is, the degree of additive Gaussian noise and INU can be altered. The "fuzzy" brain tissue images serving as "ground truth" to which the segmentation results can be compared are used to simulate the digital MRI brain phantoms.

High-resolution 3D T1-weighted volumetric MR images of digital brain phantoms using a SFLASH sequence were generated with the following parameters: TR = 15 ms, TE = 4.4 ms, and a flip angle = 25° to match closely our standard MR data acquisition protocol used in conjunction with clinical brain ^{18}F-FDG studies. Realistic brain phantoms were generated with randomly distributed noise with levels of 3%, 5%, 7%, and 9% and with INU of 0%, 20%, and 40%. Ten volumes were generated for each combination of noise and INU in the same way, only the seed is different. The latter was carefully chosen to ensure that each random sequence starts out in a different cycle, thus ensuring that the two sequences will not overlap. In addition, images with 0% noise were simulated, but with only one image per combination in this case.

Clinical data sets acquisition, registration, and delineation of brain structures

From the clinical database of the Division of Nuclear Medicine at Geneva University Hospital, eleven ^{18}F-FDG PET and five ^{18}F-DOPA studies where MR images were available were selected. The ^{18}F-FDG patients were appointed in order to detect epileptic foci with seizures, whereas the ^{18}F-DOPA patients were potentially suffering from Parkinson's disease. MR images were acquired on a Philips 1.5-Tesla Eclipse scanner (Philips Medical Systems, Best, The Netherlands) using a 3D T1-weighted gradient-echo sequence. The parameters were as follows: TR = 15 ms, TE = 4.4 ms, and a flip angle of 25°. The matrix of the images consisted of 256 × 256 × 160 voxels, with a resolution in the transaxial direction of 0.97 × 0.97

mm^2 and an axial resolution of 1.1 mm (Zaidi et al., 2003). ^{18}F-FDG PET data acquisition was performed in 3D mode on an ECAT ART (CTI/Siemens, Knoxville, TN, USA) PET tomograph upgraded to use ^{137}Cs single-photon sources to improve transmission image quality. The patients were intravenously administered approximately 222 MBq of ^{18}F-FDG. The emission study lasted 25 min and began 30 min post-injection. In order to account for photon attenuation, a 10 min pre-injection transmission scan (TX) was performed for each patient. The resulting matrix obtained after reconstruction was 128 × 128 × 47 with a resolution of 1.72 × 1.72 mm^2 in the transaxial plane and 3.38 mm in the axial direction. The scanning protocol for ^{18}F-DOPA is distinct from that of the ^{18}F-FDG PET acquisition where dynamic emission data were acquired (12 frames, 10 min/frame). The intravenously administered dose of ^{18}F-DOPA had an activity of ~185 MBq. Clinical interpretation and PVC were performed on images reconstructed over the sum of the last 6 frames (frames 7 to 12), whereas for coregistration purposes, images corresponding to the sum of all 12 dynamic frames were used. The extracted brain MR images were coregistered to their respective PET images using the automated image registration (AIR 5.2.5) software (Woods et al., 1998). A linear registration process with a traditional nine parameter model was used.

The brain MR images of those patients were automatically segmented into GM, WM, and CSF using the three algorithms introduced above. Obviously, segmentation of MRI in PET space was more challenging for the algorithms. A T1-weighted probability template and templates for the striatum (caudate nucleus and putamen) were obtained from the International Consortium for Brain Mapping (ICBM). The T1-weighted ICBM template was spatially normalized using SPM2 to the brain extracted clinical MRI of each patient realigned to the PET image, owing to the fact that the ICBM T1 template did not include the skull. The same transformation matrix was applied to the striatum template followed by converting the normalized caudate and putamen images to regions of interest (RoIs) for extracting those structures from the clinical MR images. It should be noted that one of the patients had to be excluded from this study because its striatum could not be identified due to a great loss of brain tissue. The assessment of the impact of the different MR image segmentation algorithms on PVC estimates in ^{18}F-FDG and ^{18}F-DOPA clinical PET studies was performed as described below.

Partial volume effect correction

Partial volume correction was performed using the geometric transfer matrix (GTM)-based method proposed by Rousset et al. (1998), which allows to compute corrected estimates without *a priori* knowledge on any activity level. The algorithm consists in directly computing the effect of signal degradation due to limited spatial resolution on the mean regional concentration within a limited region of space or RoI. This approach allows to obtain as many equations as there are unknowns. For instance, the observed activity t_j within tissue component D_j from a given RoI$_j$ is given by Rousset et al. (1998):

$$t_j = \sum_{i=1}^{N} \omega_{ij} T_i \tag{1}$$

where T_i represents the true tracer concentration within tissue component i. The weighting factors represent the fraction of true activity T_i from tissue i that is integrated in the measurement t_j from RoI$_j$ of volume v_j. They can be expressed as:

$$\omega_{ij} = \frac{1}{v_j} \int_{\mathrm{RoI}_j} \mathrm{RSF}_i(r)\mathrm{d}r \tag{2}$$

where RSF$_i(r)$ represents the regional spread function of tissue i and corresponds to the response of the scanner in terms of its PSF $h(r)$ to the distribution of activity D_i:

$$\mathrm{RSF}_i = \int_{D_i} h(r, r')\mathrm{d}r' \tag{3}$$

The GTM comprises a set of weighting factors ω_{ij} expressing the distortions introduced by the limited intrinsic spatial resolution of the PET scanner as well as smoothing introduced during image reconstruction and further modulation during extraction of regional tracer concentration (RoI analysis). In practice, these partial volume factors are computed from simulation of the noise-free RSF images and sampling with a user-defined set of RoIs as supplied by the segmented MRI. The set of linear equations is solved for the true tracer concentration in each region by inverting the GTM matrices and multiplying by the observed regional values (Rousset and Zaidi, 2005).

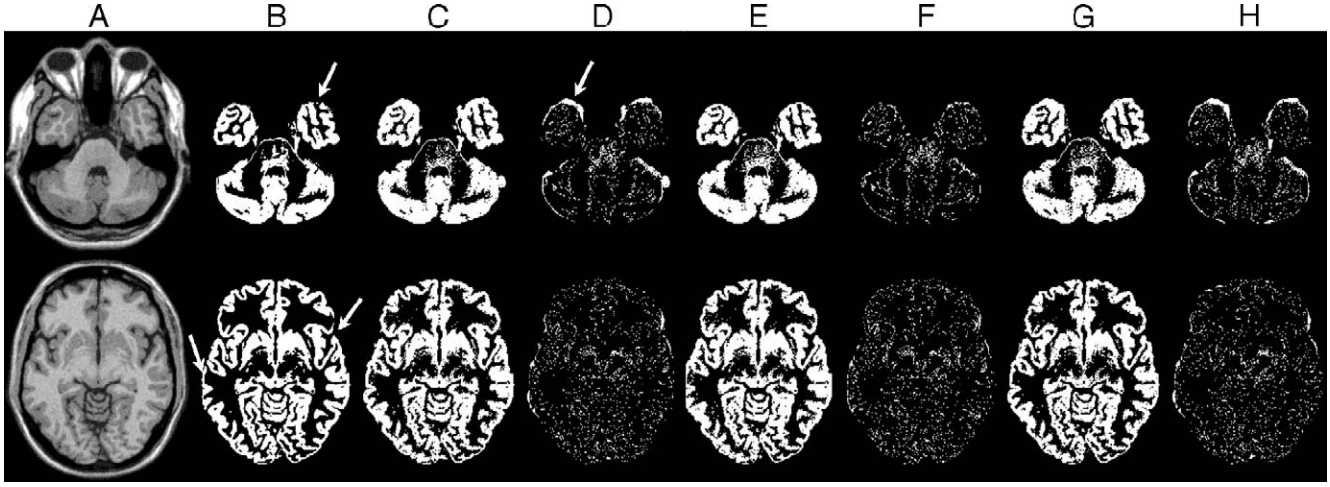

Fig. 1. Representative slices of gray matter segmentation of simulated MRI brain phantoms. A: original MRI; B: ground truth; C: SPM2; D: |B−C|; E: EMS; F: |B−E|; G: HBSA; H: |B−G|.

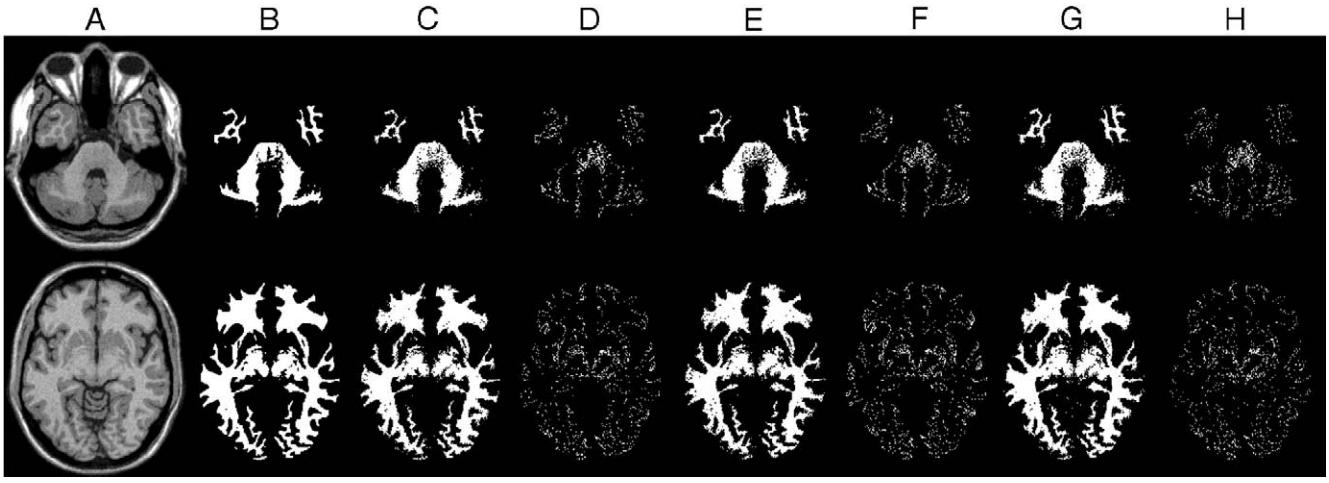

Fig. 2. Same as Fig. 1 for white matter segmentation.

Comparative assessment strategy

Quantitative assessment of the performance of different segmentation algorithms was performed by comparing the segmentation results with the "fuzzy" images of each tissue class. Error measurement assessment was performed for the EMS and SPM2 algorithms giving probability images as output (Archibald et al., 2003). First, the absolute average norm over all voxels in a given probability image A is computed as:

$$|A| = \frac{1}{N_x * N_y * N_z} \sum_{ix}^{Nx} \sum_{iy}^{Ny} \sum_{iz}^{Nz} \left| a_{ix,iy,iz} \right| \tag{4}$$

Second, the error measurement (EM) was calculated as:

$$\mathrm{EM}\left(A_{\mathrm{ref}}, A_{\mathrm{seg}}\right) = |A_{\mathrm{ref}} - A_{\mathrm{seg}}| \tag{5}$$

where A_{ref} is the "fuzzy" probability image of the investigated tissue class and A_{seg} that of the corresponding segmented tissue class. To check similarities between two images, the kappa index (κ), commonly used in reliability analysis when there is a much larger number of background voxels than that of the target voxels, was used as a figure of merit (Shattuck et al., 2001; Van Leemput et al., 1999b; Archibald et al., 2003; Grabowski et al., 2000). For two binary images (I_{ref} and I_{seg}), the kappa index is defined as:

$$\kappa\left(I_{\mathrm{ref}}, I_{\mathrm{seg}}\right) = \frac{2|I_{\mathrm{ref}} \cap I_{\mathrm{seg}}|}{|I_{\mathrm{ref}}| + |I_{\mathrm{seg}}|} \tag{6}$$

The more similar the two images are, the more the result of Eq. (6) approaches the value of unity. Moreover, to quantify the error based on mis-segmented voxels, the multiclass Type I (TI) and Type II (TII) errors were computed (Zhang, 1996; Kovacevic et al., 2002). The percentages of the two error types are given by:

$$\mathrm{TI} = 100 \times \frac{\text{Number of voxels of class } k \text{ not classified as } k}{\text{Total number of voxels of class } k} \tag{7}$$

$$\mathrm{TII} = 100 \times \frac{\text{Number of voxels of other classes than } k \text{ classified as } k}{\text{Total number of voxels of other classes than } k} \tag{8}$$

Cross-evaluation between the abovementioned parameters obtained by each of the three segmentation algorithms was

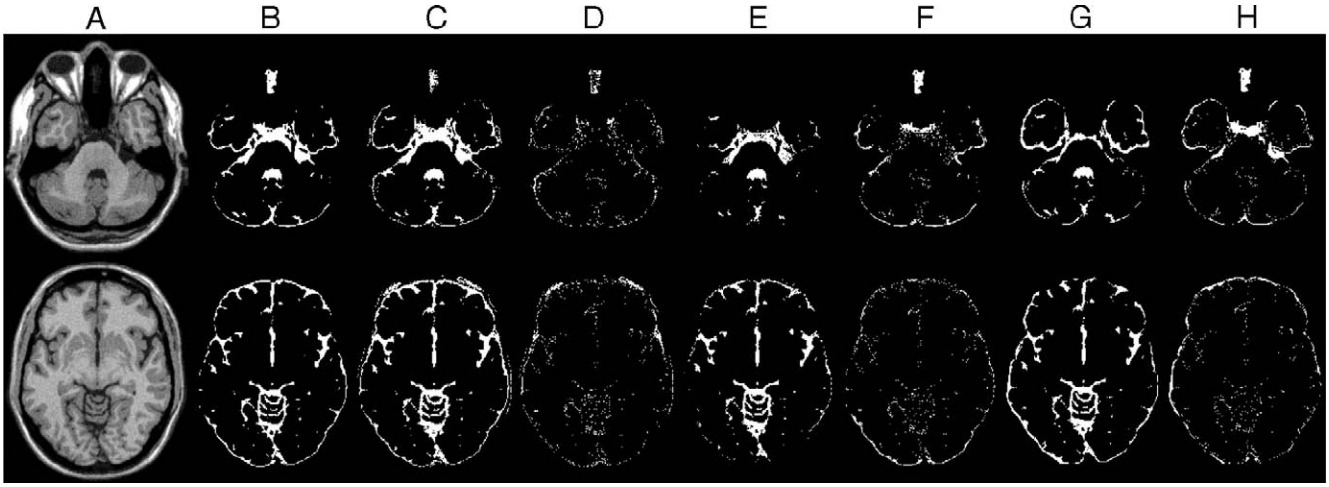

Fig. 3. Same as Fig. 1 for CSF segmentation.

performed using Student's *t* test to see whether or not one algorithm could be replaced by another for the criteria described above.

Likewise, the segmentation results of clinical MR images were compared to each other to assess potential similarities and differences between the algorithms under clinical conditions. Using tissue volumes as a figure of merit, paired correlation between two segmentation algorithms was performed by computing the linear regression for each combination of 2 segmentation algorithms. Quantification of the agreement between two segmentation algorithms was performed by using a statistical method consisting in plotting the difference against the average of the compared methods (Bland and Altman, 1995). In addition, the impact of MR image segmentation on PVC estimates in ^{18}F-FDG and ^{18}F-DOPA clinical PET studies was evaluated by passing each of the labeled segmentation results (including WM, GM, putamen, and caudate nuclei) as input to the PVC algorithm. Worth mentioning is that the authors introducing the measurement of agreement (difference vs. mean) (Bland and Altman, 1995) suggested the use of *a priori* data to which the limits of agreement should be compared. Those *a priori* data are meant to reflect the difference between two methods, which would not lead a physician to change the decision for treatment planning. Furthermore, if the degree of agreement between the two methods is not affecting patients' health, they could be used interchangeably. Unfortunately,

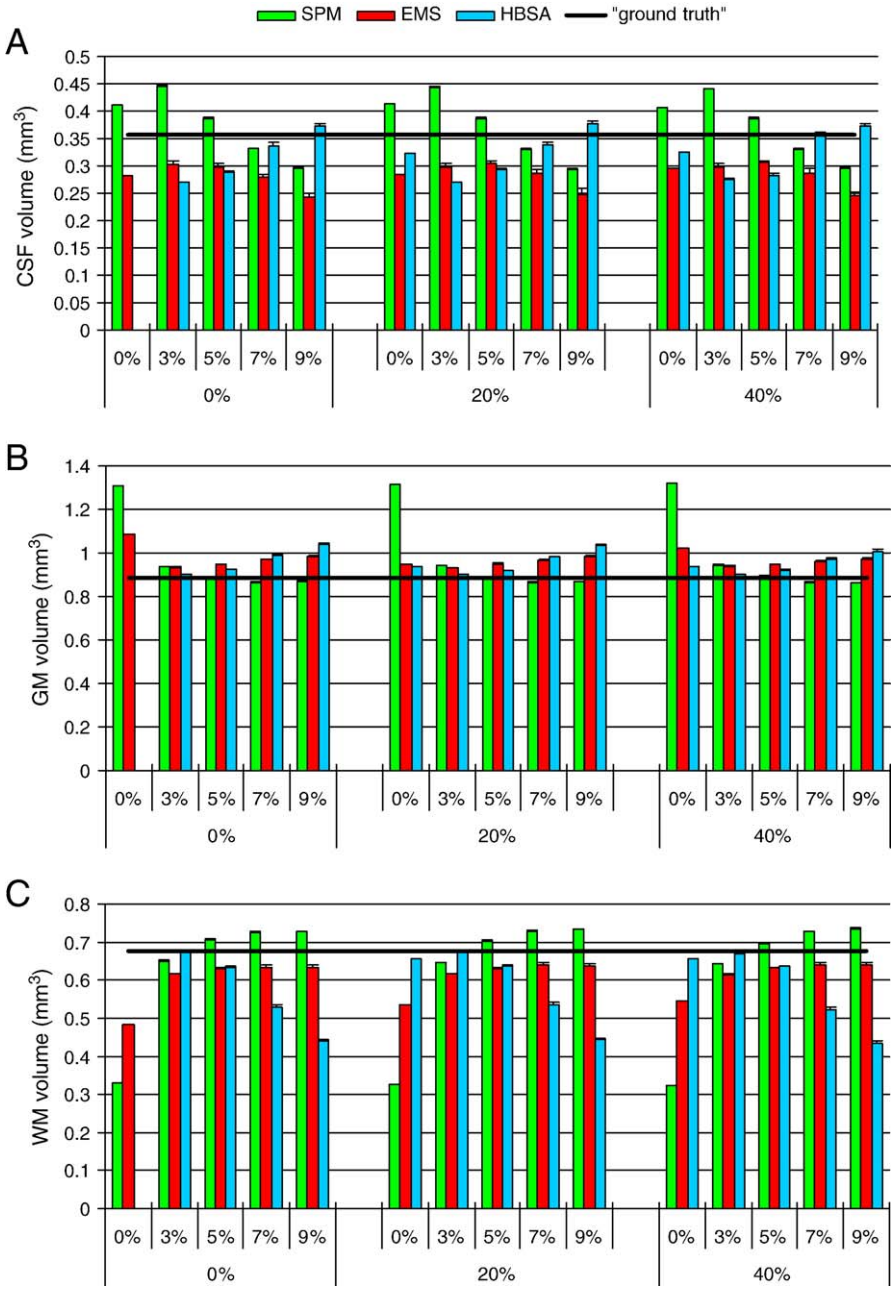

Fig. 4. Volumes assessment of segmented brain tissues. Inner values of the *x* axis correspond to the noise level, whereas the outer values correspond to the intensity non-uniformity (volumes are reported in 106 mm^3). The horizontal dark line corresponds to the binarized "ground truth" volumes estimated as 356,666 mm^3 for the CSF (A), 887,091 mm^3 for the GM (B) and 674,662 mm^3 for the WM (C).

this *a priori* knowledge is unknown in our case. However, the plots were used to provide additional information.

Results

Simulated MRI brain phantoms

Figs. 1, 2, and 3 show representative slices of the original and segmented simulated MRI brain phantoms using the algorithms described above separately for each tissue class. The noise level and the INU were set to 3% and 40%, respectively. Arrows (limited to SPM2 segmentation results) indicate possible errors in the GM fuzzy model already reported by Lemieux et al. (2003). In general, all GM segmentation results include the "glial" region. Those cells form a thin layer along the border to the CSF. It is frequently reported that this mis-segmentation occurs (Kovacevic et al., 2002; Ruan et al., 2000). Arrows point to visually apparent mis-segmentations with respect to the "ground truth". The anterior/inferior part of the temporal lobes is not correctly segmented with SPM2 and HBSA as compared to the "ground truth" in contrast to EMS which apparently segments those regions correctly. No fundamental discrepancies were observed when comparing the WM results. It was observed that non-brain matter is often included into the CSF region for both SPM2 and EMS segmentations. CSF located at the nasal level is only resolved by SPM2. Fig. 4 illustrates estimated volumes for the three tissue classes resulting from different segmentation algorithms. Tissue volumes are computed by counting the voxels belonging to each binary mask obtained by optimal thresholding and scaling by the voxel volume. Taking the relative difference as a figure of merit where the "ground truth" serves as gold standard, HBSA achieves the closest match within the 0% and 3% noise level category with regard to GM. For 3% noise, the relative difference is within the range 1.6% for 0% INU and 1.8% for 40% INU. Above 5% noise, SPM2 segments the GM volume closest to the reference, differing from −2.6% (9% noise and 40% INU) to the closest observed match at 5% noise and 0% INU (0.1% relative difference). HBSA computes the WM volume more accurately for all noise combinations up to 5%. At 3% noise and 20% INU, the relative difference is negligible

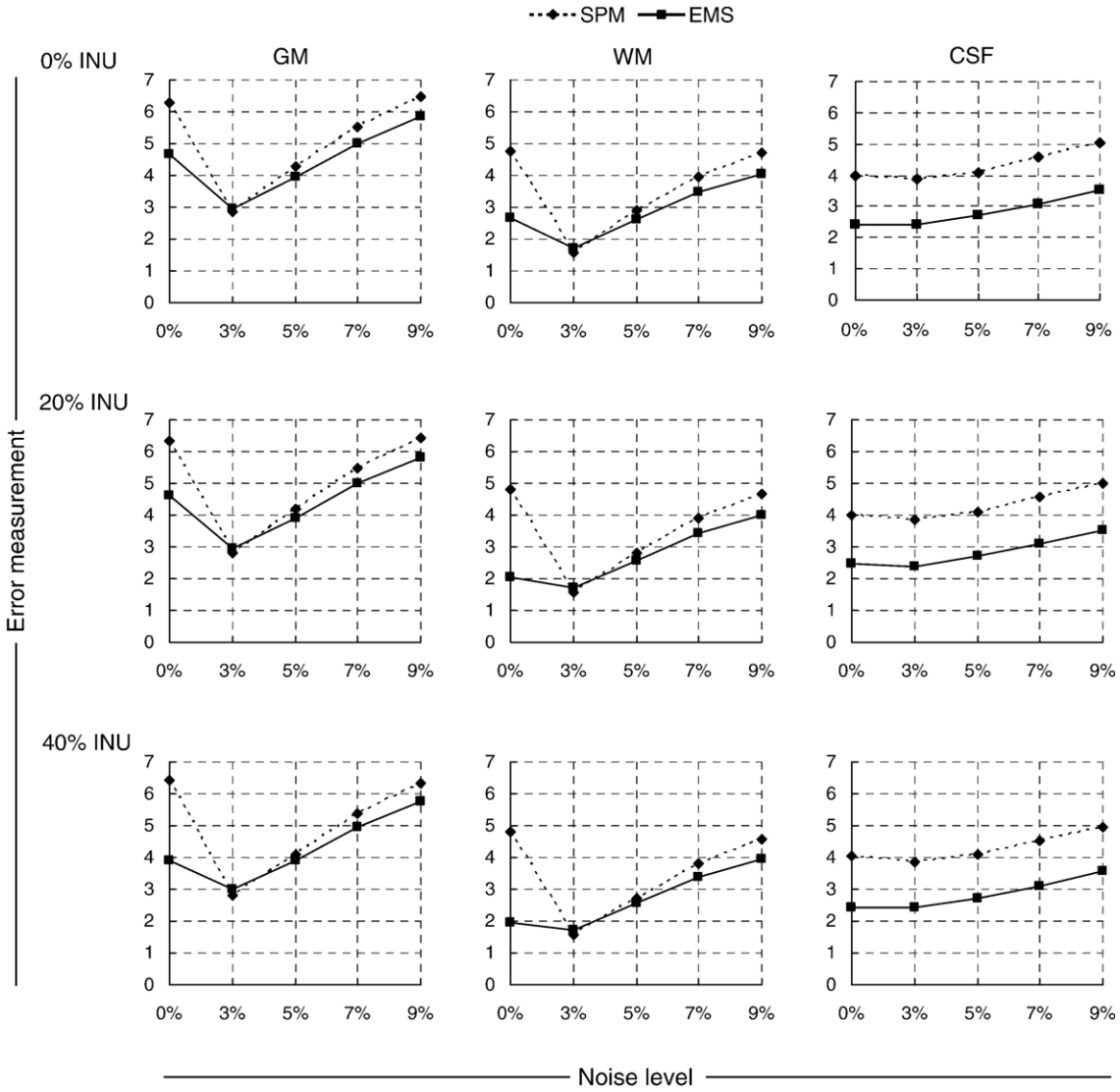

Fig. 5. Tissue probability map error measurement statistics for EMS and SPM2 segmentation algorithms using randomly generated MRI digital brain phantoms.

(0.01%). From a noise level of 7%, EMS gave the best results while HBSA peaks up to −35.4% in differing from the reference. SPM2 and the EMS remain more robust to increasing noise contamination. The HBSA segments CSF in the noise/INU combinations of 0 and 7–9% better than EMS and SPM2. When the volume is the subject of interest, EMS and SPM2 perform best at 3% and 5% noise levels for the CSF, respectively.

The probability images obtained by segmenting the digital phantoms using both SPM2 and the EMS algorithms are compared with the probability images of the "ground truth" using the error measurement as figure of merit (Fig. 5). In all 0% noise contamination combinations, the EMS gave better results. For GM, the EMS metrics ranged between 3.9 (40% INU) and 4.65 (0% INU), while SPM2 includes values between 6.28 (0% INU) and 6.42 (40% INU). However, all combinations of 3% noise are better modeled by the SPM2 software with respect to GM and WM. For a real MRI scanning environment corresponding to approximately 3% noise and 20% INU (Grabowski et al., 2000; Kovacevic et al., 2002), SPM2 culminates with 2.82, while EMS reaches 2.96 for GM classification. The WM values corresponding to that level of contamination for SPM2 are 1.56 while for EMS they equal 1.7. For all tissue classes segmented with noise combinations greater than 3%, EMS provides lower values for the error measurement. The CSF is best modeled by EMS under all conditions. In Fig. 6, for every INU condition and tissue class, the kappa metric is plotted against increasing noise contributions. Similarity between the GM and WM "ground truth" images and the respective segmented images peaked for both EMS and SPM2 algorithms at 3% noise (kappa ~0.92 for 3% noise and 20% INU). The HBSA has its highest kappa value (0.95) at almost noise-free images with either 20% or 40% INU. However, as mentioned earlier, it gave unusable results for "ideal" images.

Although all algorithms gave very similar results for the segmentation of phantoms with 3% noise for the GM, a two-tailed Student's t test revealed that EMS could be replaced by HBSA just for the combinations 3% noise and 20% or 40% INU (p = 0.44 and 0.09, respectively). As the noise level increases (>5%), the EMS gives the best results for all tissue classifica-

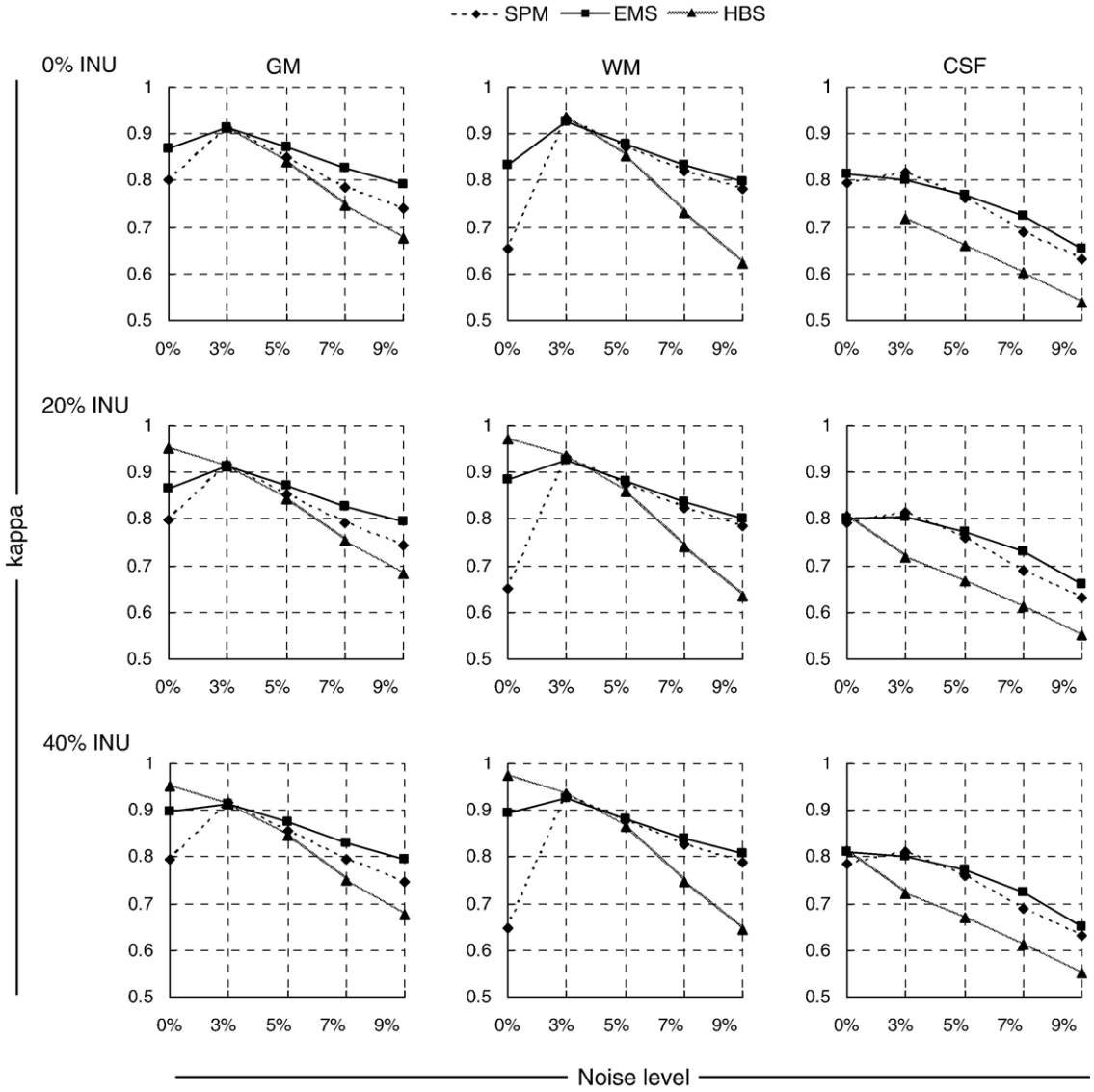

Fig. 6. Kappa metric for the 3 segmentation algorithms using randomly generated MNI digital brain phantoms.

tions. For 3/20% noise/INU combination reported to correspond to realistic contamination introduced in a clinical MRI scan, the HBSA segments the WM with the highest kappa metric (0.94). This trend is also evident for all combinations of noise/INU with a noise component lower than 3%. At higher noise and INU contributions, EMS reaches highest overlap with the reference WM image. For example, for the most extreme tested case (9% noise and 40% INU), EMS achieves a value of 0.81, while HBSA lacks behind at 0.65 and SPM2 culminates with 0.79 under these conditions. The CSF is best modeled by SPM2 package in 3% noise range with values between 0.81 (40% INU) and 0.82 (0% INU). Above 3% noise, the EMS segmentation reaches the closest similarity. At 0% noise level, HBSA and EMS gave similar results.

In Tables 1, 2, and 3, the algorithms' misclassification rates are given as the proportion of "ground truth" tissue classes. The error rates are also shown where Type I error corresponds to the probability that voxels of a certain tissue class k are misclassified as another, while Type II error expresses the probability that voxels of other tissue classes are misclassified as belonging to tissue class k. The HBSA gave unusable results when segmenting noise-free images (0% noise and 0% INU). An in-depth investigation of the reasons for the high Type I error with respect to CSF revealed that the most striking errors occur at the cerebellar level. SPM2 gives apparently the best results for the CSF segmentation (data not shown).

Clinical data

Fig. 7 shows representative slices of a clinical T1-weighted MR image, and the corresponding segmentation results when using the 3 algorithms. At anterior and inferior temporal lobe levels, SPM2 and EMS segmentation results contain more regions identified as CSF when compared to HBSA. The white arrows point to discrepancies between segmentation results. In a slice at the cerebellar level, the basal parts of the temporal and occipital lobes near the eyes were extracted as non-brain matter by the EMS. Furthermore, HBSA segmentation of the cerebellum shows some WM structures resulting in a non-uniform appearance. The HBSA misclassifies parts of the putamen regions as WM, although belonging to GM. Upper structures identified as CSF by EMS and SPM2 are not present in the HBSA results.

Table 1
Misclassified voxels and multiclass Type I and Type II error statistics for the SPM2 segmentation algorithm using simulated MRI brain phantoms

Noise	INU	True classification	Algorithm misclassification			Type I error	Type II error
			GM	WM	CSF		
3%	0%	GM	–	6.37 ± 0.08	1.22 ± 0.05	6.14 ± 0.16	5.11 ± 0.15
		WM	4.78 ± 0311	–	0.00 ± 0.00	8.45 ± 0.12	4.54 ± 0.06
		CSF	5.71 ± 0.26	0.00 ± 0.00	–	8.29 ± 0.15	0.69 ± 0.03
	20%	GM	–	6.55 ± 0.05	1.23 ± 0.04	5.68 ± 0.08	4.70 ± 0.07
		WM	4.31 ± 0.04	–	0.00 ± 0.00	8.68 ± 0.06	4.67 ± 0.04
		CSF	5.44 ± 0.19	0.00 ± 0.00	–	8.73 ± 0.08	0.70 ± 0.02
	40%	GM	–	6.68 ± 0.06	1.23 ± 0.05	5.42 ± 0.10	4.48 ± 0.08
		WM	4.04 ± 0.06	–	0.00 ± 0.00	8.85 ± 0.08	4.76 ± 0.04
		CSF	5.32 ± 0.21	0.00 ± 0.00	–	9.48 ± 0.09	0.70 ± 0.03
5%	0%	GM	–	8.20 ± 0.11	2.97 ± 0.08	15.11 ± 0.15	12.77 ± 0.13
		WM	15.16 ± 0.18	–	0.00 ± 0.00	10.89 ± 0.14	5.85 ± 0.08
		CSF	8.26 ± 0.26	0.00 ± 0.00	–	20.63 ± 0.18	1.69 ± 0.04
	20%	GM	–	8.03 ± 0.11	2.92 ± 0.05	14.55 ± 0.15	12.28 ± 0.13
		WM	14.54 ± 0.20	–	0.00 ± 0.00	10.67 ± 0.15	5.73 ± 0.08
		CSF	8.01 ± 0.18	0.00 ± 0.00	–	20.66 ± 0.15	1.66 ± 0.03
	40%	GM	–	8.09 ± 0.09	2.94 ± 0.07	13.91 ± 0.19	11.73 ± 0.16
		WM	13.86 ± 0.16	–	0.00 ± 0.00	10.75 ± 0.12	5.77 ± 0.06
		CSF	7.68 ± 0.24	0.00 ± 0.00	–	20.99 ± 0.19	1.67 ± 0.04
7%	0%	GM	–	11.21 ± 0.19	4.78 ± 0.05	22.30 ± 0.30	18.86 ± 0.26
		WM	21.97 ± 0.37	–	0.02 ± 0.00	14.93 ± 0.25	8.01 ± 0.14
		CSF	12.99 ± 0.18	0.06 ± 0.01	–	33.69 ± 0.16	2.73 ± 0.03
	20%	GM	–	10.82 ± 0.14	4.77 ± 0.07	21.98 ± 0.18	18.58 ± 0.15
		WM	21.85 ± 0.28	–	0.02 ± 0.00	14.40 ± 0.18	7.73 ± 0.10
		CSF	12.38 ± 0.21	0.05 ± 0.00	–	33.59 ± 0.23	2.72 ± 0.04
	40%	GM	–	10.60 ± 0.11	4.70 ± 0.09	21.65 ± 0.13	18.27 ± 0.11
		WM	21.50 ± 0.20	–	0.01 ± 0.00	14.09 ± 0.14	7.57 ± 0.08
		CSF	12.16 ± 0.34	0.04 ± 0.00	–	33.75 ± 0.28	2.68 ± 0.05
9%	0%	GM	–	13.82 ± 0.08	7.10 ± 0.06	26.67 ± 0.13	22.42 ± 0.11
		WM	25.97 ± 0.13	–	0.11 ± 0.00	18.66 ± 0.11	9.95 ± 0.06
		CSF	15.70 ± 0.24	0.32 ± 0.01	–	42.40 ± 0.16	4.08 ± 0.04
	20%	GM	–	13.33 ± 0.09	7.07 ± 0.09	26.57 ± 0.18	22.31 ± 0.15
		WM	26.13 ± 0.17	–	0.10 ± 0.01	17.93 ± 0.12	9.58 ± 0.07
		CSF	15.10 ± 0.29	0.26 ± 0.01	–	42.38 ± 0.24	4.06 ± 0.05
	40%	GM	–	13.06 ± 0.14	6.96 ± 0.08	26.39 ± 0.12	22.12 ± 0.10
		WM	25.93 ± 0.21	–	0.09 ± 0.00	17.52 ± 0.19	9.38 ± 0.10
		CSF	14.90 ± 0.28	0.24 ± 0.01	–	42.23 ± 0.18	3.99 ± 0.04

Misclassified voxels are expressed as the proportion of the true tissue class.

Table 2
Same as Table 1 for the EMS segmentation algorithm

Noise	INU	True classification	Algorithm misclassification			Type I error	Type II error
			GM	WM	CSF		
3%	0%	GM	–	8.53 ± 0.09	0.91 ± 0.04	6.21 ± 0.28	3.94 ± 0.23
		WM	2.88 ± 0.03	–	0.00 ± 0.00	11.45 ± 0.14	6.08 ± 0.07
		CSF	5.95 ± 0.65	0.00 ± 0.00	–	26.11 ± 1.22	0.52 ± 0.02
	20%	GM	–	8.43 ± 0.11	0.98 ± 0.08	6.34 ± 0.18	3.44 ± 0.22
		WM	2.73 ± 0.04	–	0.00 ± 0.00	11.33 ± 0.14	6.01 ± 0.08
		CSF	4.80 ± 0.64	0.00 ± 0.00	–	26.26 ± 0.93	0.56 ± 0.05
	40%	GM	–	8.63 ± 0.14	1.00 ± 0.10	6.15 ± 0.40	3.09 ± 0.28
		WM	2.55 ± 0.10	–	0.00 ± 0.00	11.59 ± 0.19	6.16 ± 0.10
		CSF	4.11 ± 0.76	0.00 ± 0.00	–	26.26 ± 0.69	0.57 ± 0.06
5%	0%	GM	–	11.30 ± 0.18	2.90 ± 0.10	9.92 ± 0.12	7.96 ± 0.09
		WM	8.55 ± 0.16	–	0.00 ± 0.00	15.13 ± 0.24	8.06 ± 0.13
		CSF	6.83 ± 0.31	0.00 ± 0.00	–	29.48 ± 1.11	1.64 ± 0.06
	20%	GM	–	11.15 ± 0.11	2.93 ± 0.11	9.51 ± 0.22	7.70 ± 0.11
		WM	8.28 ± 0.12	–	0.00 ± 0.00	14.91 ± 0.15	7.96 ± 0.08
		CSF	6.60 ± 0.31	0.00 ± 0.00	–	28.26 ± 0.72	1.66 ± 0.06
	40%	GM	–	10.86 ± 0.15	2.92 ± 0.06	9.59 ± 0.17	7.61 ± 0.07
		WM	8.28 ± 0.14	–	0.00 ± 0.00	14.52 ± 0.19	7.75 ± 0.1
		CSF	6.36 ± 0.16	0.00 ± 0.00	–	28.12 ± 0.60	1.66 ± 0.03
7%	0%	GM	–	14.53 ± 0.32	5.35 ± 0.25	13.50 ± 0.24	11.26 ± 0.12
		WM	13.09 ± 0.46	–	0.00 ± 0.00	19.36 ± 0.41	10.37 ± 0.23
		CSF	7.81 ± 0.57	0.02 ± 0.01	–	35.31 ± 1.15	3.04 ± 0.14
	20%	GM	–	13.87 ± 0.30	5.65 ± 0.39	13.57 ± 0.24	11.21 ± 0.18
		WM	13.49 ± 0.52	–	0.00 ± 0.00	18.50 ± 0.44	9.89 ± 0.22
		CSF	6.91 ± 0.57	0.01 ± 0.00	–	34.26 ± 1.81	3.21 ± 0.22
	40%	GM	–	13.64 ± 0.30	5.38 ± 0.29	13.55 ± 0.19	11.04 ± 0.22
		WM	13.16 ± 0.49	–	0.00 ± 0.00	18.20 ± 0.41	9.73 ± 0.21
		CSF	7.03 ± 0.38	0.01 ± 0.00	–	34.63 ± 1.77	3.06 ± 0.17
9%	0%	GM	–	16.93 ± 0.32	7.84 ± 0.48	16.47 ± 0.21	13.44 ± 0.17
		WM	16.24 ± 0.50	–	0.01 ± 0.00	22.57 ± 0.44	12.08 ± 0.23
		CSF	8.14 ± 0.59	0.02 ± 0.01	–	44.91 ± 1.56	4.46 ± 0.28
	20%	GM	–	16.55 ± 0.43	7.90 ± 0.60	16.44 ± 0.25	13.31 ± 0.23
		WM	16.26 ± 0.70	–	0.01 ± 0.00	22.08 ± 0.59	11.81 ± 0.31
		CSF	7.75 ± 0.68	0.02 ± 0.01	–	44.19 ± 2.22	4.49 ± 0.34
	40%	GM	–	16.07 ± 0.38	7.65 ± 0.41	16.61 ± 0.37	13.14 ± 0.16
		WM	16.08 ± 0.53	–	0.01 ± 0.00	21.46 ± 0.50	11.46 ± 0.27
		CSF	7.57 ± 0.61	0.02 ± 0.01	–	45.03 ± 1.38	4.35 ± 0.23

Table 4 shows the relative differences between segmentation algorithms with tissue volume as parameter of interest, where MRI were segmented in PET space. For the GM, a low correlation was observed between the 3 segmentation algorithms when segmented tissue volume is the measure of interest. However, the measurement of agreement demonstrates that the estimates are almost in line with the mean of the differences. In the comparative assessment considering WM results of EMS and HBSA, a low correlation was also observed. For the CSF, high correlations can be obtained for all crossover comparisons. MRI of one child was used to demonstrate the limitations of template-based segmentation and is identified since some outliers were found. Two patients were only used for SPM2/HBSA analysis since those algorithms were able to deal successfully with these data suffering from big tissue losses whereas EMS failed to register them to the template.

When comparing CSF results, a linear correlation between the differences of the volumes and the means of the volumes was apparent (data not shown). However, by disregarding problematic data mentioned above, no obvious correlation is observed. Roughly, the entire differences lie within the 95% confidence interval (limits of agreement), defined as mean ± 1.96 * SD. An outlier was identified where possible reasons might be inappropriately extracted brains when using the brain extraction utility for HBSA segmentation. Regions where some degree of atrophy was present were extracted as non-brain matter.

Impact on partial volume effect correction

PVC was carried out twice. First, the clinical MRI were segmented in PET space, and the resulting 3 segmented tissue classes were subsequently used to correct for PVE. In a second step, the PU and the CN were delineated on the MRI and used to evaluate contributions of those structures to PVE. Figs. 8 and 9 show typical results reporting correlation and agreement between recovered activity concentrations in different regions for 2 respective segmentation methods (EMS and HBSA). Identified problems were brain extraction complications (I) and conditions confusing the algorithms (II). Comparing the output of PVC with respect to the applied segmentation algorithm revealed high correlation between the results ($\geq$0.95). No obvious correlation between the mean of the difference and the average of activity values was observed with increasing values. However, similar distribution of data points in the difference vs. average plots was obtained when HBSA corrected and uncorrected values were

Table 3
Same as Table 1 for the HBSA segmentation algorithm

Noise	INU	True classification	Algorithm misclassification			Type I error	Type II error
			GM	WM	CSF		
3%	0%	GM	–	4.70 ± 0.02	1.28 ± 0.01	7.95 ± 0.05	5.91 ± 0.03
		WM	6.69 ± 0.03	–	0.00 ± 0.00	6.29 ± 0.03	3.35 ± 0.02
		CSF	4.44 ± 0.07	0.00 ± 0.00	–	37.02 ± 0.23	0.73 ± 0.01
	20%	GM	–	4.77 ± 0.02	1.25 ± 0.02	7.69 ± 0.09	5.69 ± 0.04
		WM	6.31 ± 0.03	–	0.00 ± 0.00	6.37 ± 0.03	3.40 ± 0.02
		CSF	4.51 ± 0.08	0.00 ± 0.00	–	36.90 ± 0.37	0.71 ± 0.01
	40%	GM	–	5.05 ± 0.04	1.26 ± 0.04	7.85 ± 0.15	5.75 ± 0.05
		WM	6.02 ± 0.03	–	0.00 ± 0.00	6.73 ± 0.05	3.60 ± 0.03
		CSF	5.22 ± 0.09	0.00 ± 0.00	–	36.06 ± 0.46	0.71 ± 0.02
5%	0%	GM	–	12.98 ± 0.06	2.17 ± 0.07	14.34 ± 0.10	11.40 ± 0.07
		WM	11.10 ± 0.04	–	0.00 ± 0.00	17.17 ± 0.08	9.26 ± 0.04
		CSF	11.96 ± 0.20	0.00 ± 0.00	–	40.22 ± 0.34	1.23 ± 0.04
	20%	GM	–	12.40 ± 0.10	2.04 ± 0.07	14.14 ± 0.13	11.23 ± 0.07
		WM	10.72 ± 0.12	–	0.00 ± 0.00	16.41 ± 0.14	8.85 ± 0.07
		CSF	12.19 ± 0.32	0.00 ± 0.00	–	39.39 ± 0.39	1.16 ± 0.04
	40%	GM	–	12.14 ± 0.08	2.31 ± 0.07	13.90 ± 0.51	10.19 ± 0.12
		WM	10.24 ± 0.12	–	0.00 ± 0.00	16.07 ± 0.11	8.66 ± 0.06
		CSF	10.10 ± 0.29	0.00 ± 0.00	–	40.01 ± 0.94	1.31 ± 0.04
7%	0%	GM	–	26.22 ± 0.54	3.16 ± 0.16	20.89 ± 0.38	17.06 ± 0.33
		WM	12.99 ± 0.34	–	0.00 ± 0.00	34.71 ± 0.73	18.77 ± 0.39
		CSF	24.75 ± 1.48	0.24 ± 0.03	–	415.59 ± 0.28	1.80 ± 0.09
	20%	GM	–	25.29 ± 0.50	3.01 ± 0.07	20.76 ± 0.15	16.90 ± 0.11
		WM	12.81 ± 0.28	–	0.00 ± 0.00	33.45 ± 0.67	18.09 ± 0.36
		CSF	24.64 ± 0.77	0.18 ± 0.02	–	40.22 ± 0.42	1.71 ± 0.04
	40%	GM	–	25.42 ± 0.32	2.63 ± 0.07	21.46 ± 0.33	16.82 ± 0.13
		WM	11.13 ± 0.21	–	0.00 ± 0.00	33.62 ± 0.42	18.19 ± 0.23
		CSF	27.57 ± 0.63	0.20 ± 0.01	–	38.70 ± 0.44	1.50 ± 0.04
9%	0%	GM	–	36.25 ± 0.22	4.20 ± 0.17	26.56 ± 0.42	21.64 ± 0.27
		WM	13.44 ± 0.22	–	0.02 ± 0.00	48.63 ± 0.32	26.32 ± 0.17
		CSF	37.16 ± 1.10	1.62 ± 0.10	–	44.79 ± 0.43	2.39 ± 0.10
	20%	GM	–	35.43 ± 0.41	4.14 ± 0.18	25.94 ± 0.57	20.98 ± 0.29
		WM	12.75 ± 0.31	–	0.02 ± 0.00	47.37 ± 0.56	25.64 ± 0.30
		CSF	36.53 ± 1.23	1.30 ± 0.08	–	43.39 ± 0.78	2.36 ± 0.10
	40%	GM	–	35.08 ± 0.24	3.81 ± 0.12	27.97 ± 0.55	19.88 ± 0.21
		WM	11.24 ± 0.23	–	0.01 ± 0.00	46.86 ± 0.33	25.35 ± 0.18
		CSF	36.21 ± 0.76	1.16 ± 0.04	–	43.75 ± 0.50	2.17 ± 0.07

compared either to EMS or SPM2. The distribution of data points shown in the difference vs. average plots (Fig. 9) obtained for GM and WM after separation of the PU and CN is visually identical to those assessed without separation. Furthermore, the correlation reached between results of 2 respective algorithms was excellent ($R^2 \geq 0.95$) for all regions. However, a trend is shown for the PU and CN, where activities rise or fall with increasing average activity values when considering EMS vs. HBSA and SPM2 vs. HBSA analysis. These observations cannot be deduced for SPM2 vs. EMS comparisons.

Table 5 displays the relative difference between the PVC results when comparing 2 algorithms for the clinical ^{18}F-FDG PET studies. When comparing each algorithm with respect to uncorrected vs. recovered activity values following PVC, a hierarchy was observed for the relative differences. The HBSA gave the lowest EMS medium, while SPM2 segmentation provided highest differences between GM recovered and uncorrected activities. No consistent trend was noted for the WM.

For the ^{18}F-DOPA studies, MRI was segmented in PET space and subsequently the GM was separated into GM, CN, and PU. Second, the latter labeled MRI was then furthermore delineated. That is, the left and right putamen and caudate nuclei were distinguished to investigate if increasing amounts of labels would affect PVC. Depicted in Table 6, SPM2/EMS comparisons of GM uncorrected and corrected values have the leaning to give lower relative differences than cross-evaluations where HBSA is involved. The CN corrected values follow a graduation, with lowest differences between SPM2 and EMS followed by EMS/HBSA comparisons. This hierarchy could not be clearly extracted from FDG studies. Collating the effect of HBSA and SPM2 on PVC provides largest differences. No clear conclusions can be drawn from the WM and PU values.

Discussion and conclusion

The specific role of PET imaging in the expansion of our understanding of the pathophysiological mechanisms of neurological and psychiatric diseases and in the clinical management of patients is steadily progressing. During recent years, non-invasive molecular mapping of brain function with PET has improved markedly through the development of dedicated high-resolution instrumentation and the synthesis of new ligands (Zaidi and Montandon, 2006). Despite the progress made, PET images are

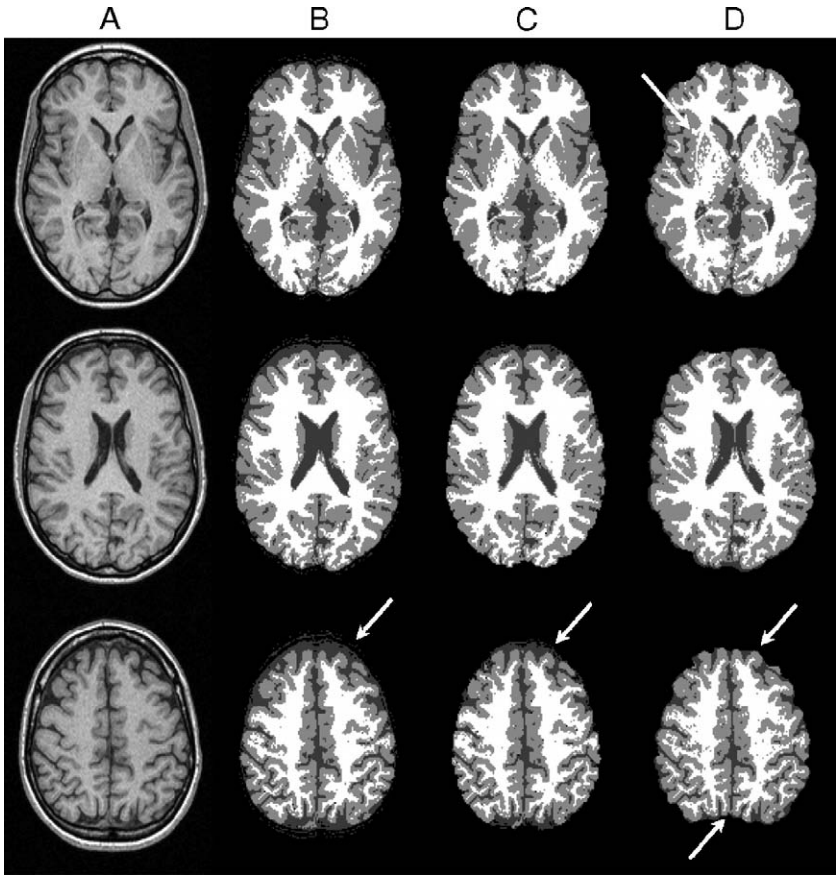

Fig. 7. Representative slices of clinical T1-weighted MR images segmentation. A: original MRI; B: SPM2; C: EMS; and D: HBSA. White corresponds to WM, light dark to GM, and dark gray to CSF. Arrows show discrepancies between segmentation results.

still suffering from many physical degrading factors (e.g., attenuation, scatter, partial volume effect) (Zaidi and Sossi, 2004). MRI-guided PVC using the GTM method is now a well-established procedure that has been confirmed by a myriad of groups and is making its way to the clinical routine (Rousset and Zaidi, 2005).

The accuracy of the PVC method depends in part on the degree of accuracy in the realignment of the anatomical MR images with the PET emission images. This has been investigated for both the pixel-based method (Muller-Gartner et al., 1992; Meltzer et al., 1996; Strul and Bendriem, 1999; Quarantelli et al., 2004) as well as for the GTM approach (Rousset et al., 1998; Frouin et al., 2002; Quarantelli et al., 2004; Slifstein et al., 2001). For the GTM approach, it is interesting to note that errors introduced during mis-registration only affect the observed estimates and do not modify the coefficients of the GTM matrix. As a consequence, the

Table 4
Relative differences (scaled to 100%) between the 3 segmentation algorithms with respect to GM and WM volume estimates of the clinical MR images after realignment to PET images

	GM			WM		
	(EMS-HBSA)/HBSA	(SPM2-EMS)/EMS	(SPM2-HBSA)/HBSA	(EMS-HBSA)/HBSA	(SPM2-EMS)/EMS	(SPM2-HBSA)/HBSA
1[a]	–		−5.28	–	–	2.80
2	10.04	1.30	8.61	8.98	1.35	7.51
3	9.73	1.97	11.89	6.91	3.19	3.94
4	1.84	6.44	8.40	23.25	4.37	17.86
5[b]	19.81	10.52	7.21	24.65	3.91	19.78
6	11.30	0.27	11.60	0.30	1.13	1.43
7	9.98	2.09	7.69	10.62	0.55	10.01
8	2.99	2.78	5.86	9.22	2.92	6.04
9[c]	12.85	0.16	12.99	15.53	4.47	11.75
10[a]	–	–	14.76	–	–	7.34
11	5.20	7.03	12.60	7.40	2.72	4.48

[a] Data used only for SPM vs. HBSA analysis.
[b] Child data.
[c] Bad brain extraction.

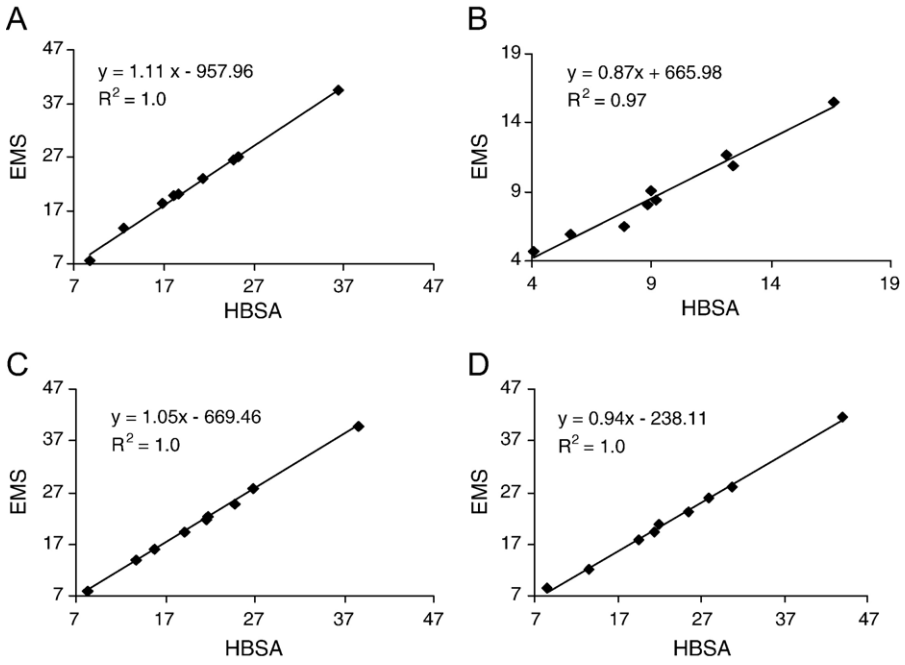

Fig. 8. Correlation plots between partial volume corrected activity values (reported in 10^3) of clinical 3D brain scans comparing results obtained using EMS (abscissa) and HBSA (ordinate) together with best fit equations and correlation coefficients for GM (A); WM (B); CN (C); and PU (D).

registration error effect on the corrected estimates is of the same magnitude as the effect of mis-registration on the observed estimates due to poor RoI placement (Slifstein et al., 2001). Those errors have been found to have relatively little impact (<2% of true value for typical 1–2 mm mis-registration error) on the final accuracy of the corrected estimates (Slifstein et al., 2001; Frouin et al., 2002). As for errors in segmentation of the tissue components of the brain, they have been found to be of greater significance with for example a 5% decrease in caudate nucleus ARC if a 25% error in total volume is made (Frouin et al., 2002). However, it has been shown that the effect of the segmentation error was limited to the mis-segmented region (Frouin et al., 2002). Overall, it appears that the success of the segmentation of the structural information provided by, e.g., MR images, has a higher impact on the accuracy of the corrected estimates compared to the influence of image coregistration, although some authors recently suggested that mis-

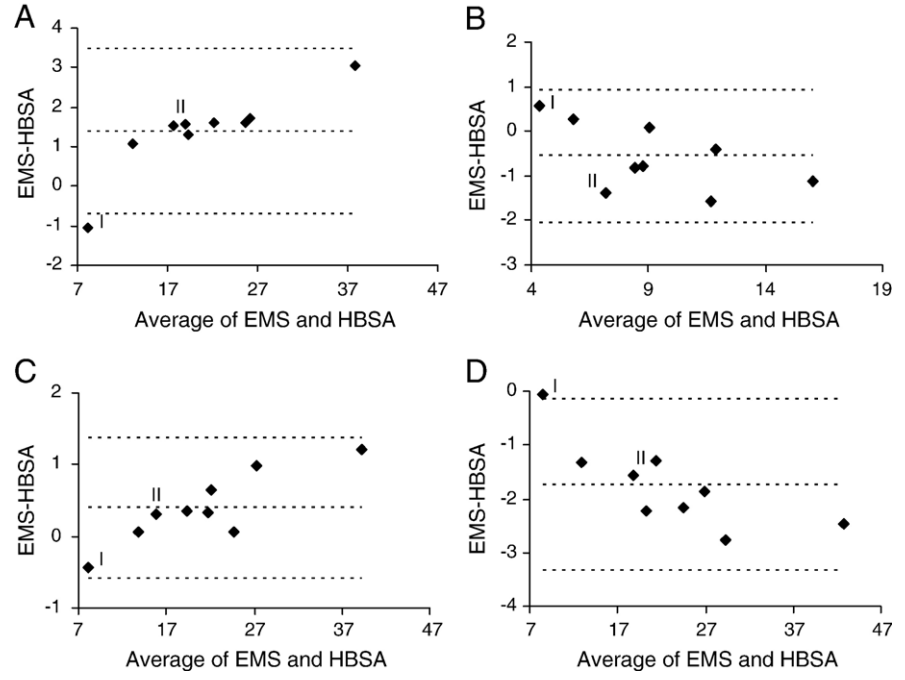

Fig. 9. Bland and Altman plots showing difference against average partial volume corrected activity values obtained using EMS and HBSA methods. The middle line is the mean, and the upper and lower broken lines are the mean ± 1.96 * SD (standard deviation). I: bad brain extraction, II: child data.

Table 5
Relative difference (scaled to 100%) between activity concentration estimates in GM and WM before and after PVC when using different segmentation algorithms on clinical ^{18}F-FDG brain PET images

	(EMS-HBSA)/HBSA	(SPM2-EMS)/EMS	(SPM2-HBSA)/HBSA
GM uncorrected			
1[a]	–	–	3.85
2	1.32	0.76	0.57
3	1.63	0.83	0.79
4	0.31	0.04	0.35
5[b]	1.49	0.49	1.01
6	0.31	0.11	0.20
7	0.12	0.02	0.15
8	0.10	0.41	0.51
9[c]	3.19	1.09	4.32
10[a]	–	–	1.54
11	0.01	0.34	0.34
GM corrected			
1	–	–	2.25
2	7.77	0.82	8.52
3	7.99	1.31	9.19
4	6.02	4.79	10.52
5	8.24	1.15	7.18
6	7.94	1.44	9.26
7	7.06	1.00	7.98
8	6.49	2.10	8.45
9	13.76	3.74	9.51
10	–	–	8.97
11	6.50	4.51	10.71
WM uncorrected			
1[a]	–	–	1.38
2	1.30	0.86	2.17
3	1.93	1.17	0.78
4	4.38	0.08	4.29
5[b]	5.83	1.01	4.76
6	1.44	1.30	0.16
7	2.33	1.01	3.37
8	2.59	0.69	3.29
9[c]	3.42	1.48	1.99
10[a]	–	–	0.69
11	0.37	0.35	0.72
WM corrected			
1	–	–	4.71
2	9.95	1.52	8.28
3	2.89	0.21	3.10
4	15.77	3.18	12.08
5	24.01	6.16	16.37
6	0.43	0.09	0.34
7	11.55	1.24	10.17
8	12.07	3.56	8.07
9	12.42	0.19	12.59
10	–	–	8.53
11	5.90	0.93	4.92

[a] Data used only for SPM vs. HBSA analysis.
[b] Child data.
[c] Bad brain extraction.

registration errors have the strongest impact on data accuracy and precision (Quarantelli et al., 2004). This recent finding is also in contradiction with the conclusions drawn for the pixel-based approach of Müller-Gärtner (Muller-Gartner et al., 1992; Strul and Bendriem, 1999). The accuracy of this latter method further depends upon the accuracy in measurement of background (i.e. WM) activity concentration. This contribution has been estimated as being in the order of 5% error in GM PET estimate for a 20% error in WM tracer concentration (Muller-Gartner et al., 1992). As for the overall performance, in the absence of major sources of registration or segmentation errors, partial volume corrected estimates have been found to be typically within 5–10% of true tracer concentration with a standard deviation of a few percent in both phantom and simulation studies (Rousset et al., 1998; Slifstein et al., 2001; Aston et al., 2002; Frouin et al., 2002; Quarantelli et al., 2004).

This study focused on the evaluation of three commonly used MRI brain segmentation algorithms. In addition, their impact on MRI-guided PVC was assessed using clinical data. Data obtained from the digital brain studies show clearly that there is no given

Table 6
Relative difference (scaled to 100%) between activity concentration estimates in different brain structures including the putamen and caudate nuclei before and after PVC when using different segmentation algorithms on clinical ^{18}F-DOPA brain PET images

	(EMS-HBSA)/HBSA	(SPM2-EMS)/EMS	(SPM2-HBSA)/HBSA
GM uncorrected			
1	5.86	4.29	9.90
2	4.75	1.83	6.49
3	8.39	3.21	11.32
4	6.60	4.11	10.43
5	2.87	0.49	2.39
GM corrected			
1	10.25	10.38	19.57
2	12.04	5.95	17.28
3	13.88	7.55	20.38
4	11.79	10.55	21.09
5	9.29	1.14	10.32
WM uncorrected			
1	0.64	0.88	1.52
2	2.68	0.08	2.59
3	0.41	0.49	0.90
4	0.59	0.98	1.58
5	0.94	0.27	1.21
WM corrected			
1	3.49	1.99	5.55
2	9.11	0.74	8.30
3	4.69	1.27	6.02
4	4.31	3.15	7.59
5	4.73	0.92	3.76
CN corrected			
1	4.61	1.20	5.75
2	8.60	1.56	10.03
3	3.97	3.59	7.42
4	6.11	2.74	8.68
5	2.35	1.22	3.54
PU corrected			
1	0.04	0.17	0.13
2	1.17	0.43	0.73
3	0.31	0.11	0.20
4	0.17	0.26	0.43
5	2.25	0.09	2.16

The differences are calculated on the basis of the first called algorithm.

segmentation algorithm providing the best results under all examined conditions. Rather, data imply an interchangeable use according to the contaminations of noise/INU of the acquired MR images. In addition, performance of the segmentation algorithms is tissue-specific, that is, an optimal combination of segmentation algorithms may be chosen according to the tissue class of interest. The accuracy of brain MR image segmentation could also be influenced by the algorithm used for brain image extraction (Boesen et al., 2004). Most important for clinical use are the contributions of noise/INU corresponding to real MR settings. It was reported that a noise level of 3% with an INU of 20% corresponds to similar (Grabowski et al., 2000) or less (Kovacevic et al., 2002) contributions of those factors in clinical MR data acquired in the authors' facility, respectively. Thus, special attention was given to the results obtained under these conditions. The HBSA showed best performance on WM classification.

Type I and Type II errors, corresponding to the probabilities of being mis-segmented as another tissue type (Type I), or that other clusters might be wrongly classified as the tissue of interest (Type II), reach lowest values (6.4% and 3.4%, respectively) for the HBSA with respect to WM under 3% noise and 20% INU contributions. Furthermore, the observed ability of HBSA to perform best on WM quantification is also supported by a striking kappa metric (0.94) and the lowest relative volume differences. The kappa statistic contains a contribution that is essentially a correction for guessing. The GM kappa metrics demonstrate similar results for EMS and SPM2 (0.92). However, a Student's t test revealed significant differences ($p < 0.05$) between both algorithms. Error measurement values suggest a better performance of SPM2 on GM, while relative volume differences are lower with the EMS (4.8%) compared to SPM2 (6.1%). However, the error measurement is a more rigorous parameter since it takes into account the spatial distribution and the voxel intensities of the output in the form of probabilistic images. Further investigation of the Type I error revealed that GM is less misclassified to another tissue when segmented with SPM2 (5.70%) rather than EMS (6.34%). The GM is less contaminated by other tissues when segmented with EMS. Finally, visual inspection of the segmented brain phantoms with respect to the GM fuzzy image favors the use of EMS. It should be noted that errors in the segmentation of the fuzzy GM images have been reported (Lemieux et al., 2003). Although the results are very similar and uncertainties remain, the use of the EMS segmentation is suggested to classify GM superior than SPM2 under 3% noise and 20% INU.

The classification of CSF is rather difficult since a fluid needs to be resolved compared to solid partitions such as GM and WM. There are large differences between the "ground truth" and all the respective segmentation results. Relative volume differences of 16.3% for the EMS, 24.4% for HBSA, and finally 24.5% for SPM2 results under assumed normal conditions are relatively high for clinical applications. Noteworthy, most research is rather focused on the segmentation of GM and WM (Archibald et al., 2003), while few reports are hardly taking the CSF into account (Lemieux et al., 2003). This is certainly due to its less significance for surgical planning, activation studies, and for PVC since apparently no specific tracer uptake is expected. It is worth emphasizing that some investigators compared the two- (one brain tissue and CSF compartments) and three-compartment (GM, WM, CSF) model for PVC in PET and concluded that the two-compartment approach is better suited for comparative PET studies, whereas the three-compartment algorithm is capable of greater accuracy for absolute quantitative measures (Meltzer et al., 1999).

With increasing contaminations of noise and INU, the EMS is suggested to yield best GM classification. This is supported by Type I and Type II error probabilities and is consistent with kappa metrics and analysis of the error measurement between the probability images of SPM2 and EMS. Additionally, also WM classification is best modeled with EMS since both kappa and analysis of the error measurement show best results for all tissue classes. Yet, Type I and Type II errors are comparable with a trend to be slightly worse than the results obtained by SPM2. Summarizing, the results of the simulated data (kappa and errors) seem to suggest that EMS outperforms in most cases SPM2 and HBSA for noise contributions larger than 3%. This is probably due to the fact that this algorithm specifically models bias field changes by multiplying the *prior* probability within the classification step in every voxel by the prior probability of class in the atlas, thus making the algorithm more robust in the case of very severe bias fields.

The balance between algorithmic complexity and the validity of results obtained are an important criterion when selecting a segmentation algorithm. Despite the fact that algorithms which make a large number of assumptions can frequently be straightforward, it is not necessarily true that complex algorithms will always perform better. The extra complexity must be used astutely and justified for the particular application at hand. Extra complexity can just as easily result in unreliability as in improved results (Thacker et al., 2004). In the clinical setting, it has become standard practice to use simplified segmentation algorithms compared to the often complex methods developed for research using MRI. Much improved high speed, low cost, and freely available open source image segmentation and processing resources are now available, and these have made it possible for research groups to design, build, and deliver highly sophisticated computational tools. For the algorithms assessed in this work, the algorithmic complexity of EMS and SPM2 is superior to HBSA.

Studies performed to assess the impact of brain MRI segmentation algorithms on PVC of clinical ^{18}F-FDG PET suggest that EMS could be used in place of SPM2 and vice versa. The highest relative differences for GM corrected values are 4.8%, while for WM corrected values a maximum of 6.2% can be observed. Relative corrected GM activity differences in combination with HBSA are up to 13.8%, while for WM values up to 24% were observed. Those relatively high differences are based on large volume discrepancies when segmenting the MRI resampled to the PET resolution after the coregistration procedure. The other source of error was discussed before and is due to the brain extraction procedure used for the MRI.

PVC recovered activity differences in the putamen show a decreasing trend with increasing mean activities, when pairs of SPM2/HBSA and EMS/HBSA were evaluated. This is not observed in SPM2/EMS evaluations. Since the volume and the location of that label are constant for all paired analysis, this tendency is introduced by the variation in volume and spatial localization of the GM and WM. This is supported by the relative volume differences where the MRI were segmented in PET space. Even when discarding the outliers, a substantial relative volume difference is apparent. Comparing only normal brains, the SPM2/EMS pair is differing by around 2% while in pairs where the HBSA was analyzed, the relative differences are about 10%. Strictly speaking, the HBSA measurement for PU is less than that of EMS or SPM2 at low values, while tending to be higher at increasing activity values. This results in a greater variance

compared to EMS and SPM2, respectively. The same occurs for CN measurement, although with an opposite leaning. Those observations support the possible use of either the EMS or SPM2 interchangeably rather than replacing one of them with HBSA. With regard to their recovery potential, HBSA gave the highest relative activity differences for the PU in ^{18}F-FDG studies, while SPM2 and EMS gave commensurable results confirming the above-stated recommendation. However, recovery potential for the PU obtained from the ^{18}F-DOPA studies demonstrates similar relative differences for all algorithms.

Likewise, the results obtained using clinical ^{18}F-DOPA studies show a relative increase in PU's activity of at least 22%. In the CN, an increase of at least 37% relative to the uncorrected values was observed. In ^{18}F-FDG studies, where GM, WM, and the basal ganglia (PU and CN) were investigated, relative differences showed an increase of at least 23% for the GM and a decline of WM activity ranging from 11% to 37%. For particular RoIs, the PU activity showed an increase of more than 12% and the CN of 23% after PVC. This is of particular interest in ^{18}F-FDG studies where epileptic foci may be detected. Without correction for PVC, the degree of epileptic activity of a specific region may be diluted. Likewise, in order to detect and stage Parkinson's disease, ^{18}F-DOPA PET scans are routinely used. Without accurately knowing the metabolic rate of dopamine turn over and its synthesis estimated by PET analysis, inappropriate treatment planning may occur.

Since the popular and most widely used GTM-based PVC method is highly dependent on accurately segmented MR images, a comparative assessment of three commonly used segmentation algorithms was performed in this work. When MRI of normal subjects is to be segmented, HBSA gives best results for WM classification whereas it is suggested to use the EMS for GM classification under normal conditions. However, segmentation performed on clinical MRI shows quite substantial differences, especially in the case of substantial tissue losses or when lesions are present. For the particular case of PVC where MR images need to be resampled to PET space, SPM2 and EMS show very similar results and may be used interchangeably. However, the normalization algorithm implemented in SPM2 is more robust for atypical brains frequently found in clinical routine. Therefore, SPM2 segmentation should be favored. The use of HBSA as implemented by its authors to be operated on high-resolution MRI needs to be optimized when the images are resampled to PET space allowing its use with confidence for PVC.

Acknowledgments

This work was supported by the Swiss National Science Foundation under grant SNSF 3152A0-102143 and the Research and Development Foundation of Geneva University Hospital under grant PRD-04-1-08. The authors gratefully thank Dr. N. Lobaugh for supplying the HBSA software and Dr. O. Rousset for supplying the GTM software and the anonymous reviewers for useful comments on the manuscript.

References

Archibald, R., Chen, K., Gelb, A., Renaut, R., 2003. Improving tissue segmentation of human brain MRI through preprocessing by the Gegenbauer reconstruction method. NeuroImage 20, 489–502.

Ashburner, J., Friston, K., 1997. Multimodal image coregistration and partitioning—A unified framework. NeuroImage 6, 209–217.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—The methods. NeuroImage 11, 805–821.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851.

Aston, J.A., Cunningham, V.J., Asselin, M.C., Hammers, A., Evans, A.C., et al., 2002. Positron emission tomography partial volume correction: estimation and algorithms. J. Cereb. Blood Flow Metab. 22, 1019–1034.

Baete, K., Nuyts, J., Laere, K.V., Van Paesschen, W., Ceyssens, S., et al., 2004. Evaluation of anatomy based reconstruction for partial volume correction in brain FDG-PET. NeuroImage 23, 305–317.

Bland, J.M., Altman, A.G., 1995. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 346, 1085–1087.

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., et al., 2004. Quantitative comparison of four brain extraction algorithms. NeuroImage 22, 1255–1261.

Boudraa, A., Zaidi, H., 2005. Image segmentation techniques in nuclear medicine imaging. In: Zaidi, H. (Ed.), Quantitative Analysis in Nuclear Medicine Imaging. Springer, New York, pp. 308–357.

Clarke, L.P., Velthuizen, R.P., Camacho, M.A., Heine, J.J., Vaidyanathan, M., et al., 1995. MRI segmentation: methods and applications. Magn. Reson. Imaging 13, 343–368.

Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., et al., 1998. Design and construction of a realistic digital brain phantom. IEEE Trans. Med. Imag. 17, 463–468.

Cuadra, M.B., Pollo, C., Bardera, A., Cuisenaire, O., Villemure, J.G., et al., 2004. Atlas-based segmentation of pathological MR brain images using a model of lesion growth. IEEE Trans. Med. Imag. 23, 1301–1314.

Cuadra, M.B., Cammoun, L., Butz, T., Cuisenaire, O., Thiran, J.P., 2005. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Trans. Med. Imag. 24, 1548–1565.

Duncan, J.S., Papademetris, X., Yang, J., Jackowski, M., Zeng, X., et al., 2004. Geometric strategies for neuroanatomic analysis from MRI. NeuroImage 23 (Suppl. 1), S34–S45.

Frouin, V., Comtat, C., Reilhac, A., Gregoire, M.-C., 2002. Correction of partial volume effect for PET striatal imaging: fast implementation and study of robustness. J. Nucl. Med. 43, 1715–1726.

Grabowski, T.J., Frank, R.J., Szumski, N.R., Brown, C.K., Damasio, H., 2000. Validation of partial tissue segmentation of single-channel magnetic resonance images of the brain. NeuroImage 12, 640–656.

Grau, V., Mewes, A.U., Alcaniz, M., Kikinis, R., Warfield, S.K., 2004. Improved watershed transform for medical image segmentation using prior information. IEEE Trans. Med. Imag. 23, 447–458.

Hoffman, E.J., Huang, S.C., Phelps, M.E., 1979. Quantitation in positron emission computed tomography: 1. Effect of object size. J. Comput. Assist. Tomogr. 3, 299–308.

ICBM (http://www.loni.ucla.edu/ICBM/).

Kessler, R.M., Ellis, J.R., Eden, M., 1984. Analysis of emission tomographic scan data: limitations imposed by resolution and background. J. Comput. Assist. Tomogr. 8, 514–522.

Kovacevic, N., Lobaugh, N.J., Bronskill, M.J., Levine, B., Feinstein, A., et al., 2002. A robust method for extraction and automatic segmentation of brain images. NeuroImage 17, 1087–1110.

Kwan, R.K.-S., Evans, A.C., Pike, G.B., 1999. MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans. Med. Imag. 18, 1085–1097.

Lemieux, L., Hammers, A., Mackinnon, T., Liu, R.S., 2003. Automatic segmentation of the brain and intracranial cerebrospinal fluid in T1-weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. Magn. Reson. Med. 49, 872–884.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997.

Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imag. 16, 187–198.

Meltzer, C.C., Zubieta, J.K., Links, J.M., Brakeman, P., Stumpf, M.J., et al., 1996. MR-based correction of brain PET measurements for heterogeneous gray matter radioactivity distribution. J. Cereb. Blood Flow Metab. 16, 650–658.

Meltzer, C.C., Kinahan, P.E., Greer, P.J., Nichols, T.E., Comtat, C., et al., 1999. Comparative evaluation of MR-based partial-volume correction schemes for PET. J. Nucl. Med. 40, 2053–2065.

Muller-Gartner, H.W., Links, J.M., Prince, J.L., Bryan, R.N., McVeigh, E., et al., 1992. Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. J. Cereb. Blood Flow Metab. 12, 571–583.

Pollo, C., Bach Cuadra, M., Cuisenaire, O., Villemure, J.G., Thiran, J.P., 2005. Segmentation of brain structures in presence of a space-occupying lesion. NeuroImage 24, 990–996.

Quarantelli, M., Berkouk, K., Prinster, A., Landeau, B., Svarer, C., et al., 2004. Integrated software for the analysis of brain PET/SPECT studies with partial-volume-effect correction. J. Nucl. Med. 45, 192–201.

Rousset, O., Zaidi, H., 2005. Correction of partial volume effects in emission tomography. In: Zaidi, H. (Ed.), Quantitative Analysis in Nuclear Medicine Imaging. Springer, New York, pp. 236–271.

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

Ruan, S., Jaggi, C., Xue, J., Fadili, J., Bloyet, D., 2000. Brain tissue classification of magnetic resonance images using partial volume modeling. IEEE Trans. Med. Imag. 19, 1179–1187.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13, 856–876.

Slifstein, M., Mawlawi, O., Laruelle, M., 2001. Partial volume effect correction: methodological consideration. In: Gjedde, A., Hansen, S.B., Knudsen, G.M., Paulson, O.B. (Eds.), Physiological Imaging of the Brain with PET. Academic Press, San Diego, pp. 67–75.

Strul, D., Bendriem, B., 1999. Robustness of anatomically guided pixel-by-pixel algorithms for partial volume effect correction in positron emission tomography. J. Cereb. Blood Flow Metab. 19, 547–559.

Suri, J.S., Singh, S., Reden, L., 2002. Computer vision and pattern recognition techniques for 2-D and 3-D MR cerebral cortical segmentation (Part I): a state-of-the-art review. Pattern Anal. Applic. 5, 46–76.

Thacker, N.A., Williamson, D.C., Pokric, M., 2004. Voxel based analysis of tissue volume from MRI data. Br. J. Radiol. 77 (2), S114–S125.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imag. 18, 885–896.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999b. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imag. 18, 897–908.

Viergever, M.A., Maintz, J.B., Niessen, W.J., Noordmans, H.J., Pluim, J.P., et al., 2001. Registration, segmentation, and visualization of multimodal brain images. Comput. Med. Imaging Graph. 25, 147–1451.

Woods, R.P., Grafton, S.T., Watson, J.D., Sicotte, N.L., Mazziotta, J.C., 1998. Automated image registration: II. Intersubject validation of linear and nonlinear models. J. Comput. Assist. Tomogr. 22, 153–165.

Zaidi, H., Montandon, M.-L., 2006. The new challenges of brain PET imaging technology. Curr. Med. Imag. Rev. 2, 3–13.

Zaidi, H., Sossi, V., 2004. Correction for image degrading factors is essential for accurate quantification of brain function using PET. Med. Phys. 31, 423–426.

Zaidi, H., Montandon, M.-L., Slosman, D.O., 2003. Magnetic resonance imaging-guided attenuation and scatter corrections in three-dimensional brain positron emission tomography. Med. Phys. 30, 937–948.

Zhang, Y.J., 1996. A survey on evaluation methods for image segmentation. Pattern Recogn. Lett. 29, 1335–1346.