Published in final edited form as:
Neuroimage. 2011 April 1; 55(3): 1091–1108. doi:10.1016/j.neuroimage.2010.12.067.

Brain MAPS: an automated, accurate and robust brain extraction technique using a template library

Kelvin K. Leung[*,a,b], **Josephine Barnes**[a], **Marc Modat**[b], **Gerard R. Ridgway**[a,b], **Jonathan W. Bartlett**[b,c], **Nick C. Fox**[a,1], and **Sébastien Ourselin**[a,b,1] **the Alzheimer's Disease Neuroimaging Initiative**

[a]Dementia Research Centre (DRC), UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK

[b]Centre for Medical Image Computing (CMIC), Department of Medical Physics and Bioengineering, University College London, WC1E 6BT, UK

[c]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

Abstract

Whole brain extraction is an important pre-processing step in neuro-image analysis. Manual or semi-automated brain delineations are labour-intensive and thus not desirable in large studies, meaning that automated techniques are preferable. The accuracy and robustness of automated methods are crucial because human expertise may be required to correct any sub-optimal results, which can be very time consuming. We compared the accuracy of four automated brain extraction methods: Brain Extraction Tool (BET), Brain Surface Extractor (BSE), Hybrid Watershed Algorithm (HWA) and a Multi-Atlas Propagation and Segmentation (MAPS) technique we have previously developed for hippocampal segmentation. The four methods were applied to extract whole brains from 682 1.5T and 157 3T T_1-weighted MR baseline images from the Alzheimer's Disease Neuroimaging Initiative database. Semi-automated brain segmentations with manual editing and checking were used as the gold-standard to compare with the results. The median Jaccard index of MAPS was higher than HWA, BET and BSE in 1.5T and 3T scans ($p < 0.05$, all tests), and the 1st-99th centile range of the Jaccard index of MAPS was smaller than HWA, BET and BSE in 1.5T and 3T scans ($p < 0.05$, all tests). HWA and MAPS were found to be best at including all brain tissues (median false negative rate ≤ 0.010% for 1.5T scans and ≤ 0.019% for 3T scans, both methods). The median Jaccard index of MAPS were similar in both 1.5T and 3T scans, whereas those of BET, BSE and HWA were higher in 1.5T scans than 3T scans ($p < 0.05$, all tests). We found that the diagnostic group had a small effect on the median Jaccard index of all four methods. In conclusion, MAPS had relatively high accuracy and low variability compared to HWA, BET and BSE in MR scans with and without atrophy.

Keywords

Automated brain extraction; skull-stripping; segmentation; MAPS; BET; BSE; HWA

[*]Corresponding author. kk.leung@ucl.ac.uk (Kelvin K. Leung).
[1]Denotes equal senior author.

1. Introduction

Whole brain extraction (or skull-stripping) refers to the process of separating brain (grey matter (GM), white matter (WM)) from non-brain (e.g. skull, scalp and dura) voxels in neuro-image data. Depending on the application, cerebrospinal fluid (CSF) spaces (ventricular and sulcal) may or may not be included in 'brain' segmentation. There is also variability in the inferior extent of the 'brain' extraction, but typically this includes brain stem and cerebellum and excludes cervical spinal cord. Accurate brain extraction is an important initial step in many image processing algorithms such as image registration, intensity normalisation, inhomogeneity correction, tissue classification, surgical planning, cortical surface reconstruction, cortical thickness estimation and brain atrophy estimation. For example, the inclusion of dura can result in an overestimation of cortical thickness (van der Kouwe et al., 2008), or add errors to regional volumes and atrophy estimates. On the other hand, missing brain tissue following brain extraction may lead to a spurious suggestion of regional or cortical atrophy and these errors cannot easily be recovered in subsequent processing steps. It should be noted that image processing algorithms may be more or less sensitive to such errors but all are undesirable.

For large multi-site natural history studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) or therapeutic trials, where thousands of MRI scans may require processing, segmentation algorithms which require large amounts of manual intervention are unfeasible. Robustness as well as accuracy of an automated brain extraction method are crucial to reduce the manual adjustment of method parameters or manual editing of unsuccessful or suboptimal automated brain segmentations, as such interventions are time consuming, and may decrease the reliability of the brain measures and potentially introduce bias to the results. Numerous automated whole brain extraction and skull-striping methods have been suggested (Smith, 2002; Lemieux et al., 1999; Ségonne et al., 2004; Hahn and Peitgen, 2000; Shattuck et al., 2001; Zhuang et al., 2006; Dale et al., 1999; Ward, 1999; Sandor and Leahy, 1997; Sadananthan et al., 2010). Studies comparing some of the most widely used automated methods (Brain Extraction Tool (BET) (Smith, 2002), 3dIntracranial (Ward, 1999), Hybrid Watershed algorithm (HWA) (Ségonne et al., 2004) and Brain Surface Extractor (BSE) (Sandor and Leahy, 1997)) with manual segmentations show that there is a range in accuracy of techniques. Similarity between the automated and manual skull-stripped brains using these methods as measured using a Jaccard index (intersection / union) ranged from 0.80 to 0.94 (Fennema-Notestine et al., 2006; Lee et al., 2003; Shattuck et al., 2009). Common areas of missing brain tissue using automated segmentation methods were found to be in the anterior frontal cortex, anterior temporal cortex, posterior occipital cortex and cerebellar areas. In two comparison studies of HWA, BET and BSE, HWA was found to be the best at including all the brain tissues, while BSE and BET were found to be the best at removing non-brain tissues (Fennema-Notestine et al., 2006; Shattuck et al., 2009).

It is important to test an image processing algorithm on as many different images as possible, e.g. images from different patient groups, scanner strengths, MR sequences and scanner manufacturers, in order to show that it can correctly segment images with different morphology, artifacts and characteristics. A key issue with brain extraction tools is their ability to perform adequately when there are varying amounts of cerebral atrophy present such as in Alzheimer's disease (AD). Table 1 gives an overview of brain extraction method comparison studies including sample sizes, diagnostic groups, scanner strengths and extraction algorithms used. The largest brain extraction method comparison study in the literature to date was carried out by Hartley et al. (2006) who compared BET and BSE with manual segmentations using the 1.5T proton-density (PD) weighted images of 296 elderly subjects (22% with dementia). Other comparison studies predominantly used healthy

subjects ranging from 20 1.5T T_1-weighted images of normal controls (Shattuck et al., 2001) to 68 1.5T and 3T T_1-weighted images of normal controls (Sadananthan et al., 2010). ADNI, which acquired MR images of hundreds of healthy subjects, AD subjects and subjects with mild cognitive impairment (MCI)) using 1.5T and 3T scanners, therefore provides an ideal dataset to test automated brain extraction methods on images with different morphology, artifacts and characteristics, and to confirm the results of the relative few studies which have compared the performance of brain extraction methods in healthy and dementia subjects.

Segmentation techniques based on multiple atlases have been applied to automatically and accurately segment various structures in the brain (Heckemann et al., 2006; Aljabar et al., 2009), including the caudate (Klein et al., 2008), hippocampus (Wolz et al., 2010; Leung et al., 2010a; Collins and Pruessner, 2010) and amygdala (Collins and Pruessner, 2010). These techniques select multiple atlases from a library of labeled images (referred to as 'template library' in this paper), and propagate the labels from different atlases to the target image after image registration. Decision or label fusion techniques are then applied to combine the labels from different atlases to create an optimal segmentation, which has been shown to be more accurate and robust than the individual segmentations (Heckemann et al., 2006; Warfield et al., 2004; Rohlfing and Maurer, 2007). This is analogous to the combination of the results from multiple classifiers in the pattern recognition field, which has been known to produce a more accurate and robust result than a single classifier (Kittler et al., 1998). In this paper, we compare the accuracy and variability of three established automated brain extraction methods (BET, BSE and HWA) and a multi-atlas propagation and segmentation (MAPS) technique we have previously developed for hippocampal segmentation (Leung et al., 2010a), using 682 1.5T and 157 3T MRI scans from the ADNI database. To the best of our knowledge, this is the largest comparison of automated brain extraction methods using multi-site 1.5T and 3T T_1-weighted MRI scans from healthy controls, mild cognitive impairment (MCI) and AD subjects. The large number of scans from different patient groups, scanner strengths, MR sequences and scanner manufacturers provided by ADNI allows us to compare the performance of automated brain extraction methods on images with very different morphology, artifacts and characteristics.

2. Methods and Materials

2.1. Method overview

In MAPS, the target image is first compared to all the atlases in a template library. Multiple best-matched atlases are then selected, and the labels in the selected atlases are propagated to the target image after image registration. Label fusion techniques are then applied to combine the labels from different atlases to create an optimal segmentation in the target image.

In the following methods sections, we describe the image data and the semi-automated whole brain segmentations that we used in the template library and used as the gold-standard for method comparison using cross-validation. Then, we provide details about MAPS, BET, BSE and HWA, and describe the parameter selection procedure for each method. We describe the approaches used to compare the accuracy and variability of the brain extraction methods.

2.2. Image data

Our image data consisted of 682 1.5T (200 controls, 338 MCI and 144 AD) and 157 3T (53 controls, 74 MCI and 30 AD) MRI scans from the the baseline time point of the ADNI database (www.loni.ucla.edu/ADNI). Table 2 shows the demographics of the subjects. Each individual was scanned with a number of sequences but for this study we only used the

baseline T_1-weighted volumetric scans. For 1.5T scans, representative imaging parameters were TR = 2300ms, TI = 1000ms, TE = 3.5ms, flip angle = 8°, field of view = 240 × 240mm and 160 sagittal 1.2mm-thick-slices and a 192 × 192 matrix yielding a voxel resolution of 1.25 × 1.25 × 1.2mm, or 180 sagittal 1.2mm-thick-slices with a 256 × 256 matrix yielding a voxel resolution of 0.94 × 0.94 × 1.2mm. For 3T scans, representative imaging parameters were TR = 2300ms, TI = 900ms, minimum full TE, flip angle = 8°, field of view = 256 × 240mm and 160 sagittal 1.2mm-thick-slices and a 256 × 256 matrix yielding a voxel resolution of 1 × 1 × 1.2mm. The full details of the ADNI MR imaging protocol are described in Jack et al. (2008), and are listed on the ADNI website (http://www.loni.ucla.edu/ADNI/Research/Cores/). Each exam underwent a quality control evaluation at the Mayo Clinic (Rochester, MN, USA). Quality control included inspection of each incoming image file for protocol compliance, clinically significant medical abnormalities, and image quality. The T_1-weighted volumetric scans that passed the quality control were processed using the standard ADNI image processing pipeline, which included post-acquisition correction of gradient warping (Jovicich et al., 2006), B1 non-uniformity correction (Narayana et al., 1988) depending on the scanner and coil type, intensity non-uniformity correction (Sled et al., 1998) and phantom based scaling correction (Gunter et al., 2006) with the geometric phantom scan having been acquired with each patient scan.

2.3. Semi-automated whole brain extraction

In this section, we describe the semi-automated whole brain extraction method that was used to create both the gold-standard brain segmentations for method comparison and the atlases in our template library in MAPS.

All the semi-automated whole brain segmentations were performed by trained expert segmentors at the Dementia Research Centre using the ‘Medical Image Display and Analysis Software’ (MIDAS) (Freeborough et al., 1997). The brain segmentation is described in Freeborough et al. (1997), but in summary: to separate the brain (grey and white matter) and non-brain voxels in the target image, a segmentor first selected two intensity thresholds representing the range of brain voxel intensities and the most inferior limits of the brain which excluded excess brainstem/spinal cord. Then, the segmentor used the erosion operation and manual editing to disconnect the brain from the skull. In order to recover eroded brain tissues, the segmentor applied the conditional dilation operation to dilate the voxels with intensity within 60% and 160% of the mean intensity of the eroded brain region. By dilating the voxels within an intensity window of the brain tissues, the conditional dilation prevented the inclusion of low intensity CSF and high intensity scalp. Furthermore, this helped to produce more consistent brain segmentations among different segmentors because the dilated region was restricted by the intensity window of the brain tissues. Lastly, the segmentor manually checked and edited the brain segmentation to include missing brain tissues and exclude non-brain tissues. The whole process took about 30 minutes on average for each brain.

The intra-class correlation coefficient for inter-rater reliability (ICC) was greater than 0.99 calculated from 11 expert segmentors delineating five subjects’ MR data. The ICC values for intra-rater reliability were all greater than 0.99 in all 11 expert segmentors, delineating five MR examinations twice.

To further estimate the intra-rater variability of the semi-automated brain extraction method, the same segmentor (S1) delineated the brains from a subset of 15 randomly chosen images (5 AD, 5 MCI and 5 controls) twice. Similarly, to assess the inter-rater variability, a different expert segmentor (S2) delineated the brains from the same subset of 15 images.

2.3.1. Statistical analysis—To assess the intra-rater reliability, the Jaccard indices for pairs of whole brain segmentations of the 15 randomly chosen images delineated by the expert segmentor S1 were calculated. To assess the inter-rater reliability, the Jaccard indices for pairs of whole brain segmentations of the 15 randomly chosen images delineated by the expert segmentors S1 and S2 were calculated.

2.4. Automated whole brain extraction

2.4.1. MAPS—Our template library consisted of the 682 1.5T MRI scans and the corresponding semi-automated brain segmentations obtained from Section 2.3. To facilitate the matching of the target image to the atlases in the template library, all the atlases were put into the same reference space by affinely registering to a subject (ADNI subject ID=021 S 0231, MCI male aged 60 with MMSE 29/30) with brain volume (1140 ml) near the mean brain volume of the whole group (1043 ml). The affine registration algorithm used in all our methods was based on maximising the normalised cross-correlation between the source and target images (Lemieux et al., 1994) using a conjugate gradient descent optimization scheme. Since the semi-automated brain segmentations in the template library were also used as the gold-standard for the method comparison, all experiments were performed in a leave-one-out fashion. We excluded the image being segmented from the template library, meaning that the template library effectively consisted of 681 scans for the leave-one-out experiments.

To extract the whole brain from the target image, we performed the following three steps (also see Fig. 1):

1. Template selection: the target image was affinely registered to the subject to which all the template library scans were registered. Best matches from the template library were ranked as to their similarity using the cross-correlation (R^2) between the target image and the template library over the 2-voxel dilated whole brain segmentations. Cross correlation has been shown to provide a good criterion for template selection in multi-centre imaging data (Aljabar et al., 2009). Once a rank of best to worst matches was established, a subset of the highest ranking matches could be used to propagate the undilated whole brain segmentation onto the target image.
2. Label propagation: the best-matched atlases were registered to the target image using affine registration and non-rigid registration based on free form deformation (Rueckert et al., 1999; Modat et al., 2010). Multiple control point spacings (16mm→8mm→4mm) were used in the non-rigid registration to model increasingly local deformations. The whole brain segmentations in the best-matched atlases were then propagated to the target image using the results of the registrations. The grey-level whole brain segmentation in the target image was thresholded between 60% and 160% of the mean intensity of the segmentation, followed by a 2-voxel conditional dilation within 60% and 160% of the mean intensity of the segmentation. The same intensity thresholding and 2-voxel conditional dilation was previously used to recover missing brain tissues in the automated segmentation of whole brain regions in the repeat images using the propagation of the semi-automated whole brain regions in the baseline images (Evans et al., 2009; Leung et al., 2010b).
3. Label fusion: Multiple brain segmentations in the target image were combined using label fusion. The fused segmentation was further unconditionally dilated by 2 voxels to recover any missing brain tissues because it was felt better to possibly include more non-brain tissues, than to exclude real brain tissues, as described in Ségonne et al. (2004). We referred to the dilated fused segmentation as the

automated whole brain segmentation from MAPS and the undilated one as 'undilated MAPS-brain'.

2.4.2. BET in FMRIB Software Library version 4.1.4 (http://www.fmrib.ox.ac.uk/fsl/)—BET estimates the minimum and maximum intensity values of the brain image, and evolves a deformable model to fit the brain surface based on smoothness criteria and a local intensity threshold (Smith, 2002).

2.4.3. BSE in BrainSuite version 09e (http://brainsuite.usc.edu/)—BSE uses a 2D Marr-Hildreth operator for brain edge detection after anisotropic diffusion filtering (Shattuck et al., 2001). Mathematical morphology is then used to extract the brain from the edge map.

2.4.4. HWA in FreeSurfer version 4.5 (http://surfer.nmr.mgh.harvard.edu/)—Similar to Shattuck et al. (2009), HWA combines watershed algorithms and deformable surface models (Ségonne et al., 2004). The watershed algorithm provides a robust initial estimate of the brain volume for the deformable model to fit a smooth surface around the brain. A statistical atlas is used to validate and correct the brain extraction.

2.5. Parameter selection

2.5.1. Training datasets—Our previous experiences with MAPS suggested that a relatively small number of images were sufficient to choose the reasonable parameters for the wider dataset. We randomly selected 10 1.5T scans as the training dataset for MAPS. For BET, BSE and HWA, we randomly selected 18 scans by choosing one scan from each diagnostic group (controls, MCI and AD) in each field strength (1.5T and 3T) from each scanner manufacturer (GE, Philips and Siemens), in order to provide a variety of different images in the training dataset. The best parameters were determined by comparing the results with the semi-automated brain segmentations. The best parameters were then used for our whole dataset. Note that we decided to use a larger and more evenly distributed training dataset for BET, BSE and HWA than MAPS, in order to be able to get the best possible results from them.

2.5.2. MAPS—We applied MAPS to the 10 randomly chosen 1.5T scans in order to determine the number of best-matched atlases and the optimal label fusion technique required to produce accurate 'undilated MAPS-brains' by comparing them to the semi-automated brain segmentations. We combined segmentations from 3 to 29 best-matched atlases using either voting (Heckemann et al., 2006), shape based averaging (SBA) (Rohlfing and Maurer, 2007) or simultaneous truth and performance level estimation (STAPLE) (Warfield et al., 2004). For SBA, we used the 50% trimmed mean (Rothenberg et al., 1964) instead of the simple mean when calculating the average distance of a voxel to the labels, in order to increase the robustness to outliers.

2.5.3. BET—We chose to investigate the fractional intensity threshold option '-f' (default=0.5) and the following additional mutually exclusive options: '-R' for robust brain centre estimation, '-S' for eye and optic nerve cleanup and '-B' for bias field and neck cleanup. We applied BET to the 18 randomly chosen scans using either with no option, '-R', '-S' or '-B' to determine the best mutually exclusive option. Our previous experiences with BET showed that it had a tendency to exclude some brain voxels in the results. As the documentation of BET states that a smaller fractional intensity threshold returns a larger brain region, we varied the fractional intensity thresholds between 0.0 to 0.5 (increment of 0.1) after determining the best mutually exclusive options ('-R', '-S' or '-B').

2.5.4. BSE—We chose to examine the following parameters: '-n' for the number of diffusion iterations, '-d' for the diffusion constant and '-s' for the edge constant. We applied BSE to the same 18 randomly chosen scans (used for parameter selection in BET) using the option '-p' (for post-processing dilation of the final brain mask) and all the combinations of the following parameters: '-n'=(4, 5, 6, 7, 8, 9, 10), '-d'=(14, 15, 16, 17, 18, 19, 20, 21, 22), '-s'=(0.5, 0.6, 0.7, 0.8, 0.9).

2.5.5. HWA—We chose to investigate the following parameters as Shattuck et al. (2009): '-atlas': use the atlas information to correct the segmentation, 'less': shrink the surface and 'more': expand the surface. We applied HWA to the same 18 randomly chosen scans (used for parameter selection in BET) using the following options: default, '-less', '-more', '-less -atlas' and '-more -atlas'.

2.6. Method comparison

2.6.1. Quantitative evaluation metrics—The automated whole brain segmentations were compared to the semi-automated whole brain segmentations obtained (described in Section 2.3) using the Jaccard index, false positive rate and false negative rate (Shattuck et al., 2009; Sadananthan et al., 2010):

- Jaccard index was used to measure the overlap similarity of two segmentations and is defined as $\frac{|A \cap B|}{|A \cup B|}$, where A is the set of voxels in the automated region and B is the set of voxels in the gold-standard region;
- False positive rate was used to measure the probability of false brain voxels in the automated segmentation, and is defined as $\frac{|FP|}{|TN+FP|}$, where FP is the set of false positive voxels and TN is the set of true negative voxels. It is related to the specificity by: specificity = 1 − (false positive rate);
- False negative rate was used to measure the probability of missing brain voxels in the automated segmentation, and is defined as $\frac{|FN|}{|TP+FN|}$, where FN is the set of false negative voxels and TP is the set of true positive voxels. It is related to the sensitivity by: sensitivity = 1 − (false negative rate).

Different automated brain extraction methods generated segmentations containing different amounts of CSF voxels. In order to avoid the influence of different amounts of CSF voxels included in the segmentations, we followed the comparison methods suggested by Boesen et al. (2004) and Sadananthan et al. (2010) when calculating the Jaccard index and false positive rate. Low intensity voxels were excluded from all the whole brain segmentations by using a consistent threshold. We chose the threshold as 60% of the mean intensity of the gold-standard semi-automated brain segmentation. The Jaccard index and false positive rate were then calculated using the thresholded whole brain segmentations. The false negative rate was calculated using the unthresholded whole brain segmentations.

Since the 'undilated MAPS-brains' were derived from the semi-automated whole brain segmentations, we also performed a direct comparison between them using the Jaccard index, false positive rate and false negative rate without excluding low intensity voxels. This direct comparison was not performed for BET, BSE and HWA because of the different amounts of CSF included in BET, BSE, HWA, and the 'gold-standard' semi-automated segmentations, which would make the results less meaningful.

2.6.2. Qualitative analysis using projection maps—In order to visualise the locations of the segmentation errors in different automated whole brain extraction methods, we generated projection maps of the false positive and negative voxels (Shattuck et al., 2009). All the images in our dataset were non-rigidly registered to the subject (ADNI subject ID=021 S 0231) to which all the template library scans were registered. Multiple control point spacings (16mm→8mm→4mm) were used in the non-rigid registration to model increasingly local deformations. We then affinely registered the subjects to the MNI 305 atlas (Mazziotta et al., 1995). Using the affine and non-rigid transformations, we mapped the false positive and negative voxels of all the segmentations into the MNI 305 atlas using nearest-neighbour interpolation. For each transformed false positive and negative map, we computed 2D sagittal, coronal and axial projections by summing the counts of voxels along the respective directions. Each pixel in these 2D projection maps denoted the number of erroneous voxels along a projected ray in the particular direction. To summarise all the false positive (or negative) projection maps of a brain extraction method, we calculated an average projection map from the projection maps of all the segmentations by taking the mean value of all the projection maps at each pixel.

2.6.3. Application of 'undilated MAPS-brains' in brain atrophy estimation—The boundary shift integral (BSI) provides a precise measurement of brain atrophy from two serial MR scans (Freeborough and Fox, 1997). The first step in BSI requires the extraction of the brain regions that includes GM and WM and excludes internal and external CSF from the two serial MR scans. KN-BSI was recently proposed to produce a more robust atrophy estimation in multi-site data by incorporating better intensity normalisation and automatic parameter selection (Leung et al., 2010b). We therefore compared the use of semi-automated segmentations and 'undilated MAPS-brains' in brain atrophy estimation of the baseline and 12-month 1.5T scans of our ADNI dataset using KN-BSI.

We applied MAPS to obtain 'undilated MAPS-brains' of the baseline and 12-month 1.5T scans, and used them to calculate KN-BSI (referred to as MAPS KN-BSI). We also calculated a KN-BSI using the semi-automated segmentations in the baseline scans and propagated brain segmentations in the 12-month scans as Leung et al. (2010b) and Evans et al. (2009) (referred to as semi-automated MAPS KN-BSI). The propagated brain segmentations in the 12-month scans were calculated by propagating the semi-automated segmentation from the baseline scans to the 12-month scans of *the same subject* using affine registration and nonrigid registration based on B-splines (Rueckert et al., 1999).

2.7. Statistical analysis

We compared the Jaccard index, false positive rate and false negative rate between the brain extraction methods in 1.5T and 3T scans. Due to the highly skewed distribution of the Jaccard index, false positive rate and false negative rate, the median was used to measure the average accuracy of a method, and the 1st to 99th centile range (CR) was used to measure the variability in accuracy of a method. Confidence intervals (CI) for the differences in the median and CR were found using bias-corrected and accelerated (BCa) bootstrap CIs (Efron and Tibshirani, 1993) (10,000 bootstrap samples), using STATA's bootstrap command. This procedure created 10,000 samples by sampling subjects (and their data) from the original dataset (with replacement). Since the distribution of differences was non-normal, we report whether $p < 0.05$ on the basis of whether the BCa bootstrap CI for the differences includes the null value of 0. We also performed the same analysis to assess differences in the median and CR of the Jaccard index, false positive rate and false negative rate between subject diagnostic groups and between scanner field strength within each method, which are given in the supplementary material.

We refer to an automated whole brain segmentation as 'failed' when its Jaccard index was 0, meaning that there was no overlap between the automated and semi-automated whole brain segmentations.

A pairwise *t*-test was used to compare the differences between semi-automated KN-BSI and MAPS KN-BSI in each diagnostic group. The agreement between the two KN-BSIs was further examined using a Bland-Altman plot (Bland and Altman, 1986).

3. Results

3.1. Semi-automated whole brain extraction

The mean (SD) Jaccard index between the two different semi-automated segmentations by the same segmentor S1 were 0.988 (0.005) (see Table 3(a)), and the mean (SD) Jaccard index between the different semi-automated segmentations delineated by the expert segmentors S1 and S2 were 0.989 (0.003) (see Table 3(b)). Furthermore, based on the 15 images (5 controls, 5 MCI and 5 AD), we found that the mean (SD) number of voxels modified by the expert segmentor S1 after the thresholding procedure was 6403 (3964).

3.2. Parameter selection of MAPS, BET, BSE and HWA

Fig. 2 shows the accuracy of the 'undilated MAPS-brain' using different numbers of best-matched atlases and label fusion techniques. SBA performed better than voting and STAPLE, and the accuracy of SBA started to reach a plateau when combining more than 19 segmentations. As a trade-off between accuracy and running-time, we decided to choose 19 best-matched atlases and combined them using SBA, which gave an average Jaccard index of 0.980 in the subset of 10 images. Fig. 3 demonstrates MAPS by showing the intermediate and final results using the chosen parameters.

Table 4 shows the accuracy of BET, BSE and HWA using different parameters. For BET, the best parameters were '-B -f 0.3', which gave an average Jaccard index of 0.953. For BSE, the best parameters were '-n 4 -d 20 -s 0.70 -p', which gave an average Jaccard index of 0.917. Furthermore, for HWA, the best parameters were '-less', which gave an average Jaccard index of 0.956.

3.3. Comparison of MAPS, BET, BSE and HWA

Typical performance of automated brain extraction methods in 1.5T and 3T scans in our dataset are shown in Figs. 4 and 6. In addition, Figs. 5 and 7 show examples of thresholded segmentations using 60% of the mean intensity of the semi-automated segmentation in 1.5T and 3T scans. Tables 5 and 6 show the median and CR (1st-99th centile range) of the Jaccard index, false positive rate and false negative rate of MAPS, BET, BSE and HWA using the 1.5T and 3T scans respectively. MAPS had the highest median Jaccard index, and BSE had the lowest median false positive rate. HWA, closely followed by MAPS, had the lowest median false negative rate. Furthermore, MAPS had the smallest CR in the Jaccard index, false positive rate and false negative rate. We found that while no MAPS and HWA segmentations failed, 2 BET segmentations (2 1.5T images) and 3 BSE segmentations (2 1.5T and 1 3T images) failed (see Fig. S.1(a) and S.1(b) in the supplementary material for two examples).

3.3.1. Qualitative analysis using projection maps—Non-brain tissue was included in all automated segmentation algorithms (see Fig. 8). All algorithms erroneously added dura surrounding the cerebellum (including tentorium) and cortex (including falx cerebri). Inclusion of these extra tissues appeared relatively more pronounced and extensive using HWA particularly in the tentorium and nervous tissue running medial to the temporal lobes

including optic nerves. Neck and other non-brain tissues inferior to the brain area were included in some segmentations of BET. Our false negative maps (see Fig. 9) show more discrepancies across techniques compared with the false positive maps. It is important to note the differences in scale bar when comparing across these techniques; the scale bar for MAPS and HWA extend only to 0.6 whereas BET and BSE extend to 10. Very few areas were erroneously excluded by MAPS and these areas appear to fall largely outside of the brain (for example, tentorial tissue) and may therefore represent subtle manual missegmentations (see Fig. 10). BET appeared to wrongly exclude cerebellar and occipital lobe tissue as well as anterior temporal and frontal lobe areas in some cases. The fact that the whole of the brain was visible using BET was due to complete failure of the technique in a very small number of images as described above. BSE appeared to falsely exclude cerebellar and inferior temporal lobe tissue on a number of scans. HWA, much like BSE, had some problems correctly including cerebellar tissue on some images, and in a very small number of cases (see scale bar) this extended to the remainder of the brain.

3.3.2. Between-method comparison—Tables 7 and 8 shows differences in median and CR (1st-99th centile range) of the Jaccard index, false positive rate and false negative rate between MAPS, BET, BSE and HWA.

- *Accuracy*

  There was evidence of differences in the median Jaccard index among all the automated brain extraction methods except between HWA and BET. In both 1.5T and 3T segmentations, the median Jaccard index of MAPS was higher than HWA and BET, which in turn was higher than BSE.

  There was evidence that the median false positive rates differed among all the methods. The methods in ascending order of the median false positive rate were BSE, MAPS, BET and HWA in 1.5T segmentations and BSE, BET, MAPS and HWA in 3T segmentations.

  There was evidence that all false negative rates differed among the methods except in 1.5T segmentations between HWA and MAPS. In 1.5T segmentations, the median false negative rates of MAPS and HWA were lower than BET, which in turn was lower than BSE. In 3T segmentations, the methods in ascending order of the median false negative rate were HWA, MAPS, BET and BSE.

- *Variability in accuracy*

  There was evidence of differences in the CRs of the Jaccard index among all the automated brain extraction methods except in 3T segmentations between BET, BSE and HWA. In 1.5T segmentations, the methods in the ascending order of CR of the Jaccard index were MAPS, HWA, BSE and BET. In 3T segmentations, the CR of the Jaccard index of MAPS was smaller than BET, BSE and HWA.

  There was evidence of differences in the CRs of the false positive rate among all the automated brain extraction methods except in 3T between HWA and BET. In 1.5T segmentations, the methods in ascending order of the CR of the false positive rate were MAPS, HWA, BSE and BET. In 3T segmentations, the CR of the false positive rate of BSE was smaller than MAPS, which in turn was smaller than HWA and BET.

  There was evidence of differences in the CRs of the false negative rate among all the automated brain extraction methods except in 3T between HWA, BET

and BSE. In 1.5T segmentations, the methods in ascending order of the CR of the false negative rate were MAPS, HWA, BSE and BET. In 3T segmentations, the CR of the false negative rate of MAPS was smaller than BET, BSE and HWA.

3.4. Computation time

The computation time of BSE and HWA were about 1 minute per image running on a personal computer with a Intel(R) Xeon(R) CPU (X5472 3.00GHz) and 4Gb of RAM, whereas the computation time of BET was about 10 minutes per image. The computation time of MAPS was about 19 hours because of the computationally expensive non-rigid registrations.

3.5. Direct comparison of 'undilated MAPS-brains' with semi-automated segmentations

Table 9 shows the direct comparison between the 'undilated MAPS-brains' and semi-automated segmentations. The median Jaccard index (CR) was 0.980 (0.053) and 0.974 (0.106) in 1.5T and 3T segmentations. Note that the median Jaccard index and false positive rate of 'undilated MAPS-brains' are similar to *thresholded* MAPS segmentations in Table 5. This was due to the fact that the thresholding removed most of the lower intensity voxels (e.g. CSF) after the 2-voxel dilation. On the other hand, since the false negative rate was calculated using the unthresholded MAPS segmentation, the false negative rate of the MAPS segmentation was lower than the 'undilated MAPS-brain'.

3.6. Application of 'undilated MAPS-brains' in brain atrophy estimation

We found excellent agreement between semi-automated KN-BSI and MAPS KN-BSI (see Table 10 and Fig 11), although there were small statistically significant differences between them (with semi-automated KN-BSI > MAPS KN-BSI).

3.7. Post-hoc analysis

Since our results showed that the median accuracy of MAPS was higher than BET, BSE and HWA in the ADNI dataset when using our semi-automated brain segmentations as the gold-standard, we used the Segmentation Validation Engine (SVE) website (http://sve.loni.ucla.edu/archive/) to further test MAPS on a different dataset (40 healthy subjects; mean (SD) age = 29.2 (6.3)), and compared the results with the gold-standard brain masks delineated using a different manual segmentation protocol as described in Shattuck et al. (2009). Since the brain masks provided by the SVE website included all the internal ventricular CSF and some external sulcal CSF, we slightly modified the MAPS algorithm to include them in the brain segmentation (see Appendix A for more details). The median (CR) Jaccard index of MAPS was 0.955 (0.019) (ID=173, http://sve.loni.ucla.edu/archive/study/?id=173), which was the highest amongst all the entries at the time of writing (other entries included BSE, BET, HWA, statistical parametric mapping (SPM) (Ashburner and Friston, 2005) and various other algorithms). The median Jaccard index of MAPS was 0.002 (95% CI (−0.001, 0.004), $p > 0.05$) higher than the second highest entry (which used the voxel-based morphometry (VBM) toolbox (version 8, http://dbm.neuro.uni-jena.de/vbm8/VBM8-Manual.pdf)), and the CR of the Jaccard index of MAPS was 0.009 (95% CI (−0.005, 0.013), $p > 0.05$) lower than VBM. The CIs suggested that both tests were close to statistical significance.

4. Conclusions and Discussion

We wished to evaluate a template-based automated brain extraction method (MAPS) and a number of well-established automated brain extraction methods relative to a conventional

semi-automated method that involves time consuming manual editing. We applied the four automated brain extraction methods (MAPS, BET, BSE and HWA) to over 800 scans from the ADNI database. This set of images included scans with a range of anatomy and atrophy: from healthy elderly subjects with little atrophy to MCI and AD subjects with very significant atrophy.

All four methods showed reasonable overlap (Jaccard index) with the semi-automated 'gold-standard' segmentation. Among the four methods, MAPS had higher median accuracy and smaller variability in accuracy. Both MAPS and HWA had low false negative and false positive rates, meaning that they were able to preserve nearly all the brain voxels and, at the same time, removed most of the non-brain voxels. MAPS removed more non-brain voxels than HWA and was less variable than HWA in terms of the CR of false positive rate and false negative rate. Although the median accuracy of BET was higher than BSE, the variability in accuracy of BSE was lower than BET. Of note, in the direct comparison, 'undilated MAPS-brains' were found to be very accurate, with a median Jaccard index of 0.980 in 1.5T segmentations. This is close to the mean Jaccard index of two different segmentations produced by the same segmentor (0.988) and segmentations performed by different segmentors (0.989). Furthermore, MAPS KN-BSI was in excellent agreement with semi-automated KN-BSI, and the small mean (SD) difference of 0.02% (0.08%) between them was less than the mean (SD) difference of 0.05% (0.47%) in BSI between same-day scan pairs reported by Boyes et al. (2006) in a different study.

We compared the four automated brain extraction methods qualitatively using the false positive and false negative projection maps (see Figures 8 and 9). While the false positive projection maps appear quite similar with added dura surrounding the cerebellum, the false negative projection maps show that different methods failed to include tissues in different locations as represented by different 'hot spots'. BET appeared to tend to exclude temporal and frontal lobe tissues (consistent with the findings of Shattuck et al. (2009)) as well as cerebellar tissue. Both BSE and HWA appeared to erroneously exclude cerebellar tissue. However, Shattuck et al. (2009) did not find that HWA excluded much cerebellar tissue, which was likely due to the difference in the range of morphology and characteristics of the brain images in the datasets. The results of the quantitative comparison between BET, BSE and HWA are similar to those reported by Fennema-Notestine et al. (2006), Shattuck et al. (2009) and Sadananthan et al. (2010), with HWA being better at preserving brain voxels than BET and BSE, and BET and BSE being better at removing non-brain voxels than HWA.

Although the effect of scanner field strength on the accuracy of MAPS and HWA was minimal, the effect on the robustness of HWA was large: the CR of the false negative rate in 3T segmentations is 39 percentage points higher than 1.5T segmentations. The median Jaccard index and false negative rate of BET and BSE in 1.5T segmentations were better than 3T segmentations. Although there was no evidence of a difference in the variability in the Jaccard index of BET and BSE between 1.5T and 3T segmentations, the CR of the false negative rate of BSE in 3T segmentations is 40 percentage points higher than 1.5T segmentations. Sadananthan et al. (2010) also found that the performance of the methods were different in their 1.5T and 3T datasets.

Despite the efforts put into trying to ensure that the characteristics of MR images in the ADNI dataset were similar across different scanner manufacturers and field strengths, there are inevitably significant differences and it is interesting that field strength significantly affected the accuracy and robustness of the automated brain extraction methods.

The effect of the diagnostic groups on the automated brain extraction methods was complicated; the accuracy of MAPS in all the groups was similar, however, MAPS produced slightly less robust results in controls. This is likely due to the 2-voxel dilation performed at the end of the processing as the dilated brain region in controls is more likely included non-brain tissues (e.g. dura) than MCI or AD subjects. BET produced more accurate results in controls with higher median Jaccard index and lower median false negative rate. On the other hand, there was little suggestion of the robustness of BET being different across diagnostic groups except at 3T the segmentations of AD subjects were more robust than control and MCI subjects. Although there was no evidence of a difference in the accuracy of BSE between diagnostic groups, it was surprising that the robustness of BSE was significantly better in MCI subjects in 1.5T segmentations. The accuracy of HWA in all the diagnostic groups was similar. Although there was no evidence of a difference in the robustness of HWA between diagnostic groups, the CR of the false positive rate of controls tended to be smaller than AD and MCI subjects.

Although we did not find any significant difference in the median Jaccard index of BSE and HWA between diagnostic groups, we found that BET produced significantly more accurate results in controls than MCI and AD subjects in both 1.5T and 3T scans. This was similar to the findings of Fennema-Notestine et al. (2006) that the average Jaccard index of BET in young normal controls was higher than AD subjects (Figure 5 of (Fennema-Notestine et al., 2006)).

We previously found that STAPLE was the best method to combine multiple hippocampal segmentations in terms of the Jaccard index (Leung et al., 2010a). However, we found shape based averaging to be better for whole brain segmentations. The best label fusion method is likely to be problem specific, consistent with the findings of Artaechevarria et al. (2009); in that depending on the characteristics of the images and regions, globally or locally weighted voting produced substantially better results than simple majority voting. It is interesting to note that the chosen parameters give similar results in the small subset and our whole dataset, meaning that the 10 randomly chosen 1.5T images have provided a good sample for parameter selection in MAPS. Given the excellent results in the 3T scans and the scans from SVE, the chosen parameters may also be suitable for scans acquired using different MR sequences and scanners - this potential generalisabilty (based on the range of anatomy included in the template library) is a possible advantage over those methods that require parameter selection based on a subset of scans. The oscillation in the accuracy of SBA in Fig. 2 may appear concerning in terms of performance, however it is due to the discreteness in 50% trimmed mean: the 50% trimmed mean discards equal or unequal numbers of segmentations from either side depending on the number of segmentations.

For large studies and clinical trials, it is more important to minimise the human interaction time and expertise required to correct any sub-optimal segmentation (e.g. parameter fine-tuning or manual editing) than to minimise the computation time of the algorithm. Although the computation time of MAPS is comparatively much longer than BET, BSE and HWA, the robustness of MAPS was substantially higher than the other methods. Furthermore, the processing time of MAPS can be improved by (1) running the software using a computer cluster, (2) using fewer atlases in a trade off between accuracy and computation time, or (3) running the non-rigid registration on a graphical processing unit (GPU) (Modat et al., 2010).

One of the strengths of this study is the large number of images of AD, MCI and control subjects acquired from scanners of different field strength and manufacturers at multiple sites. To the best of our knowledge, this is the largest comparison of automated brain extraction methods in the literature. Another strength of this study is that all the data and softwares will be openly available to the public on the world wide web. All the scans can be

downloaded from the ADNI website (http://www.adni-info.org). The semi-automated brain segmentations will be available on the ADNI website. BET, BSE and HWA are all available on the web (see Section 2). The registration software and label fusion softwares used in MAPS can be downloaded at http://sourceforge.net/projects/niftyreg/ and http://www.itk.org/. We will make all the MAPS brain regions available on-line at the ADNI website (http://adni.loni.ucla.edu/).[3]

One of the limitations of this study is the lack of ground-truth whole brain segmentations in the method comparison. Instead, we used semi-automated segmentations which were then manually edited by trained expert segmentors. The segmentors followed a pre-defined segmentation protocol to ensure low intra- and inter-rater variability. Another limitation is that the amount of brain stem labelled as brain may not be consistent between the semi-automated and automated segmentations. Although the thresholding was designed to remove CSF from the automated segmentations to allow the comparison with semi-automated segmentations, it may remove some grey matter from the brains and lose some important information at the boundary of the brain. We also did not try to use other label fusion algorithms in MAPS (apart from vote, SBA and STAPLE), such as a local weighted voting method (Artaechevarria et al., 2009) or a selective and iterative method (Langerak et al., 2010). In addition, although we examined most of the parameters in BET, BSE and HWA using a subset of scans from our dataset, an expert user may be able to fine-tune other parameters or use a different subset to produce better results.

Despite the fact that all the MAPS experiments were carried out in a leave-one-out fashion, MAPS may have an advantage over other methods in the comparison because the definition of a brain region in the MAPS segmentations is likely to be more consistent with the semi-automated segmentations. Partly our motivation for developing and assessing MAPS was to replace the semi-automated segmentation - there is therefore some potential intrinsic advantage to MAPS (relative to BET, BSE and HWA). As such we must be cautious about the conclusions. Nonetheless the advantage is arguably minimal because of the following:

1. The post-hoc analysis (Section 3.7) showed that MAPS performed well both in terms of accuracy and variability in accuracy on a different and independent dataset with gold-standard brain masks delineated using a different manual segmentation protocol (SVE). The comparison using SVE is not only independent but also involves a wide range of algorithms with parameters that have been fine-tuned either by the developers or Shattuck et al. (2009). Currently, SVE contains 118 sets of results from several algorithms (e.g. VBM8, BSE and brainwash2). We found that the evaluations using our semi-automated brain segmentations and the independent gold-standard segmentations from SVE are consistent with each other;
2. The final step in MAPS involved a 2-voxel unconditional dilation. Although this step was designed to recover missing brain tissues, it also substantially reduces the similarity between the MAPS segmentations and the gold-standard segmentations. For example, using a randomly chosen brain segmentation in our template library, a 2-voxel dilation reduces the Jaccard index from 1 to 0.741;
3. There is a substantially amount of manual intervention in the semi-automated segmentation, which includes the selection of the initial intensity thresholds and the editing of brain/non-brain tissues during various stages of the semi-automated segmentation;

[3]Please contact the corresponding author if you cannot locate the MAPS brain regions on the ADNI website.

4. In order to reduce the influence of the amount of CSF included in the automated brain segmentations in the comparison, the Jaccard index and the false positive rate were calculated using thresholded brain segmentations as in Sadananthan et al. (2010) and Boesen et al. (2004). The thresholding values were given by 60% of the mean brain intensity of the gold-standard segmentation. This thresholding step ensures consistent cut-off points between CSF and GM interface in all the automated segmentations;

5. The false positive rate and false negative rate maps of MAPS show errors near the inferior brain stem. This suggests that there is still inconsistency between the MAPS brain segmentations and gold-standard segmentations.

The outputs of different brain extraction algorithms include different amount of internal ventricular and external sulcal CSF. Therefore, we chose to use a consistent threshold to exclude low intensity voxels from all the brain segmentations, as suggested by Boesen et al. (2004) and Sadananthan et al. (2010), to try to compare different algorithms in as unbiased manner as possible. However, we acknowledge that brain extraction is rarely used in isolation and that dependent on the subsequent processing steps and ultimate outcome measure being assessed the quality of segmentation and possible errors included may or may not be important. The requirement for accuracy in brain extraction therefore varies with different uses of the masks. We also acknowledge that each of the other methods might well be fine-tuned to particular scan types and applications. Although we showed that the semi-automated KN-BSI and MAPS KN-BSI were very similar, future work should examine the suitability of a particular brain extraction method for the specific processing pipeline or application for which it is to be used.

In conclusion, our results suggest that a template library approach (MAPS) is a relatively accurate and robust method of automated brain extraction. MAPS was similar to HWA in the ability to preserve brain tissues, but removed significantly more non-brain tissues than HWA. MAPS was shown to be more robust than HWA. We suggest that fully automated brain extraction methods now approach the accuracy and reliability of time consuming manual techniques and may be particularly valuable in large scale studies. Ultimately, the development and evaluation of accurate and robust brain segmentation methods that are able to equal or outperform more labour-intensive manual segmentation procedures will facilitate more efficient research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 5-year public-private partnership. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The Principal Investigator is Michael W. Weiner, M.D., VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research – approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical

Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

This work was undertaken at UCLH/UCL who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The Dementia Research Centre is an Alzheimer's Research Trust Co-ordinating Centre and has also received equipment funded by the Alzheimer's Research Trust. KKL is supported by a Technology Strategy Board grant (TP1638A) working in partnership with IXICO Ltd. on the project 'Imaging to assess efficacy and safety of new treatments for Alzheimer's Diseases', NCF is funded by the Medical Research Council (UK), and JB is supported by an Alzheimer's Research Trust Research Fellowship.

The authors would like to thank all the image analysts and the research associates in the Dementia Research Centre for their help in the study. In particular, we would like to thank Raivo Kittus and Melanie Blair for performing brain segmentations for the evaluation of intra- and inter-rater variability. The implementations of voting, SBA, and hole-filling algorithms used the Insight Segmentation and Registration Toolkit (ITK), an open source software developed as an initiative of the U.S. National Library of Medicine and available at www.itk.org. We thank Simon Warfield for kindly providing us with the source code of STAPLE. The research of STAPLE was supported in part by NIH R01 RR021885 from the National Center For Research Resources, and by an award from the Neuroscience Blueprint I/C through R01 EB008015 from the National Institute of Biomedical Imaging and Bioengineering. The authors would particularly like to thank the ADNI study subjects and investigators for their participation.

References

Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage. 2009 Jul; 46(3):726–738. [PubMed: 19245840]

Artaechevarria X, Munoz-Barrutia A, de Solorzano CO. Combination strategies in multiatlas image segmentation: application to brain MR data. IEEE Trans Med Imaging. 2009 Aug; 28(8):1266–1277. [PubMed: 19228554]

Ashburner J, Friston KJ. Unified segmentation. NeuroImage. 2005 Jul; 26(3):839–851. [PubMed: 15955494]

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb; 1(8476):307–310. [PubMed: 2868172]

Boesen K, Rehm K, Schaper K, Stoltzner S, Woods R, Lders E, Rottenberg D. Quantitative comparison of four brain extraction algorithms. NeuroImage. 2004 Jul; 22(3):1255–1261. [PubMed: 15219597]

Boyes RG, Rueckert D, Aljabar P, Whitwell J, Schott JM, Hill DLG, Fox NC. Cerebral atrophy measurements using Jacobian integration: comparison with the boundary shift integral. NeuroImage. 2006 Aug; 32(1):159–169. [PubMed: 16675272]

Collins DL, Pruessner JC. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. NeuroImage. 2010 May.

Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage. 1999 Feb; 9(2):179–194. [PubMed: 9931268]

Efron, B.; Tibshirani, R. An introduction to the bootstrap. Chapman and Hall; 1993.

Evans M, Barnes J, Nielsen C, Kim L, Clegg S, Blair M, Leung K, Douiri A, Boyes R, Ourselin S, Fox N. the Alzheimer's Disease Neuroimaging Initiative. Volume changes in Alzheimer's disease and mild cognitive impairment: cognitive associations. Eur Radiol. 2009 Sep.

Fennema-Notestine C, Ozyurt IB, Clark CP, Morris S, Bischoff-Grethe A, Bondi MW, Jernigan TL, Fischl B, Segonne F, Shattuck DW, Leahy RM, Rex DE, Toga AW, Zou KH, Brown GG. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum Brain Mapp. 2006 Feb; 27(2):99–113. [PubMed: 15986433]

Freeborough P, Fox N. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Transactions in Medical Imaging. 1997; 16(5):623–629.

Freeborough PA, Fox NC, Kitney RI. Interactive algorithms for the segmentation and quantitation of 3-D MRI brain scans. Comput Methods Programs Biomed. 1997 May; 53(1):15–25. [PubMed: 9113464]

Gunter, JL.; Bernstein, MA.; Borowski, BJ.; Felmlee, JP.; Blezek, DJ.; Mallozzi, RP.; Levy, JR.; Schuff, N.; Jack, CR. ISMRM. 2006. Validation Testing of the MRI Calibration Phantom for the Alzheimer's Disease Neuroimaging Initiative Study; p. 2652

Hahn H, Peitgen H-O. The Skull Stripping Problem in MRI Solved by a Single 3D Watershed Transform. 2000

Hartley SW, Scher AI, Korf ESC, White LR, Launer LJ. Analysis and validation of automated skull stripping tools: a validation study based on 296 MR images from the Honolulu Asia aging study. NeuroImage. 2006 May; 30(4):1179–1186. [PubMed: 16376107]

Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage. 2006 Oct; 33(1): 115–126. [PubMed: 16860573]

Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell JL, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DLG, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging. 2008 Apr; 27(4):685–691. [PubMed: 18302232]

Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, Macfall J, Fischl B, Dale A. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage. 2006 Apr; 30(2):436–443. [PubMed: 16300968]

Kittler J, Hatef M, Duin RP, Matas J. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20:226–239.

Klein S, van der Heide UA, Lips IM, van Vulpen M, Staring M, Pluim JPW. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med Phys. 2008 Apr; 35(4):1407–1417. [PubMed: 18491536]

Langerak T, van der Heide U, Kotte A, Viergever M, van Vulpen M, Pluim J. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). IEEE Trans Med Imaging. 2010 Jul.

Lee J-M, Yoon U, Nam S-H, Kim J-H, Kim I-Y, Kim SI. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. Comput Biol Med. 2003 Nov; 33(6):495–507. [PubMed: 12878233]

Lemieux L, Hagemann G, Krakow K, Woermann FG. Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magn Reson Med. 1999 Jul; 42(1): 127–135. [PubMed: 10398958]

Lemieux L, Jagoe R, Fish DR, Kitchen ND, Thomas DG. A patient-to-computedtomography image registration method based on digitally reconstructed radiographs. Med Phys. 1994 Nov; 21(11): 1749–1760. [PubMed: 7891637]

Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, Schuff N, Fox NC, Ourselin S. Alzheimer's Disease Neuroimaging Initiative. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage. 2010a Jul; 51(4):1345–1359. [PubMed: 20230901]

Leung KK, Clarkson MJ, Bartlett JW, Clegg S, Jack CR, Weiner MW, Fox NC, Ourselin S. Alzheimer's Disease Neuroimaging Initiative. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. NeuroImage. 2010b Apr; 50(2):516–523. [PubMed: 20034579]

Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). NeuroImage. 1995 Jun; 2(2):89–101. [PubMed: 9343592]

Modat M, Ridgway GR, Taylor ZA, Lehmann M, Barnes J, Hawkes DJ, Fox NC, Ourselin S. Fast free-form deformation using graphics processing units. Comput Methods Programs Biomed. 2010 Jun; 98(3):278–284. [PubMed: 19818524]

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am. 2005 Nov; 15(4):869–877. [PubMed: 16443497]

Narayana P, Brey W, Kulkarni M, Sievenpiper C. Compensation for surface coil sensitivity variation in magnetic resonance imaging. Magn Reson Imaging. 1988; 6(3):271–274. [PubMed: 3398733]

Park JG, Lee C. Skull stripping based on region growing for magnetic resonance brain images. Neuroimage. 2009 Oct; 47(4):1394–1407. URL http://dx.doi.org/10.1016/j.neuroimage.2009.04.047. [PubMed: 19389477]

Rohlfing T, Maurer CR. Shape-based averaging. IEEE Trans Image Process. 2007 Jan; 16(1):153–161. [PubMed: 17283774]

Rothenberg TJ, Fisher FM, Tilanus CB. A Note on Estimation from a Cauchy Sample. Journal of the American Statistical Association. 1964; 59(306):460–463.

Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging. 1999 Aug; 18(8):712–721. [PubMed: 10534053]

Sadananthan SA, Zheng W, Chee MWL, Zagorodnov V. Skull stripping using graph cuts. NeuroImage. 2010 Jan; 49(1):225–239. [PubMed: 19732839]

Sandor S, Leahy R. Surface-based labeling of cortical anatomy using a deformable atlas. IEEE Trans Med Imaging. 1997 Feb; 16(1):41–54. [PubMed: 9050407]

Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B. A hybrid approach to the skull stripping problem in MRI. NeuroImage. 2004 Jul; 22(3):1060–1075. [PubMed: 15219578]

Shattuck DW, Prasad G, Mirza M, Narr KL, Toga AW. Online resource for validation of brain segmentation methods. NeuroImage. 2009 Apr; 45(2):431–439. [PubMed: 19073267]

Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. NeuroImage. 2001 May; 13(5):856–876. [PubMed: 11304082]

Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. 1998 Feb; 17(1):87–97. [PubMed: 9617910]

Smith SM. Fast robust automated brain extraction. Hum Brain Mapp. 2002 Nov; 17(3):143–155. [PubMed: 12391568]

van der Kouwe AJW, Benner T, Salat DH, Fischl B. Brain morphometry with multiecho MPRAGE. NeuroImage. 2008 Apr; 40(2):559–569. [PubMed: 18242102]

Ward, B. 3dintracranial: Automatic segmentation of intracranial region. 1999. URL http://afni.nimh.nih.gov/afni/doc/manual/3dIntracranial

Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004 Jul; 23(7):903–921. [PubMed: 15250643]

Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D, Initiative ADN. LEAP: learning embeddings for atlas propagation. NeuroImage. 2010 Jan; 49(2):1316–1325. [PubMed: 19815080]

Zhuang AH, Valentino DJ, Toga AW. Skull-stripping magnetic resonance brain images using a model-based level set. NeuroImage. 2006 Aug; 32(1):79–92. [PubMed: 16697666]

Appendix

A. Modified MAPS for the Segmentation Validation Engine

This section describes the modified MAPS algorithm that generated the brain regions for the Segmentation Validation Engine (SVE) (ID=173, http://sve.loni.ucla.edu/archive/study/?id=173). Since the manual brain segmentations provided by SVE include internal ventricular CSF and some external sulcal CSF, we slightly modified MAPS in Section 2.4.1 to include them in the brain segmentation. We used the same template library that consisted of 682 1.5T MRI scans. In addition to the semi-automated brain segmentations, we also used the semi-automated ventricles segmentations delineated by the trained expert segmentors at the Dementia Research Centre.

1. Intensity non-uniformity correction: the intensity non-uniformity in the target image was corrected by applying N3 (Sled et al., 1998).

   Template selection: there was no change to this step.

2. Label propagation: in order to include internal CSF, we propagated the semi-automated ventricles segmentations from the atlases to the target image, and added it to the conditionally dilated brain regions at the end of this step.

3. Label fusion: there was no change to this step. However, we used the ‘undilated MAPS-brain’ as the input to the next step.

4. Hole filling: in order to fill in any internal cavities and gaps in the ‘undilated MAPS-brain’, an iterative voting-based hole-filling image filter was applied to fill in any voxels whose $5 \times 5 \times 5$ (full width) neighbourhood had more than 64 brain voxels. The number of iterations of the hole-filling image filter was set to 5. Any remaining holes were filled by flood-filling the image background from the edge and taking the unflooded voxels as the brain region. The brain region was further dilated by 1-voxel to include some external CSF.

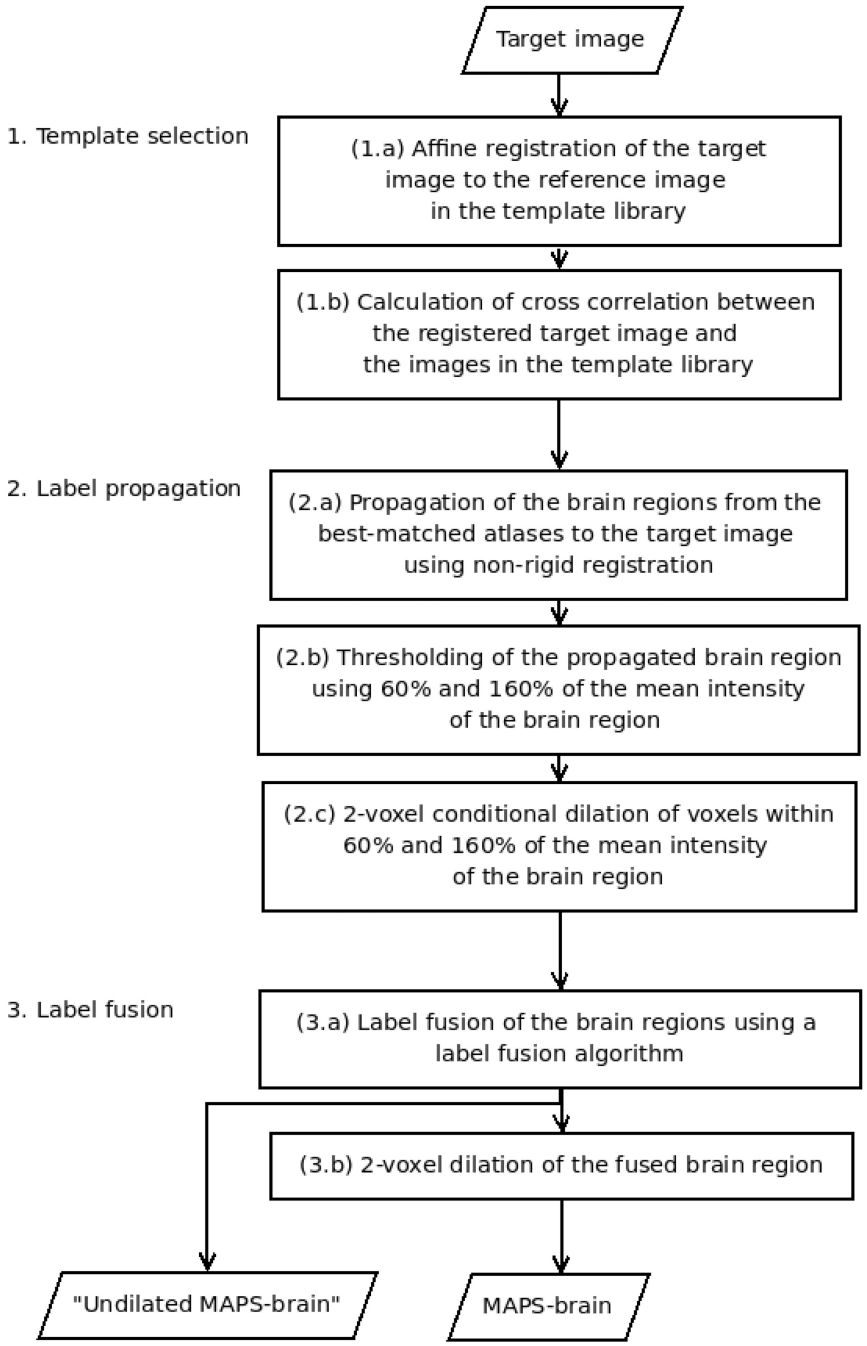


Figure 1.
The flowchart of MAPS. Please refer section 2.4.1 for the description of each processing step.

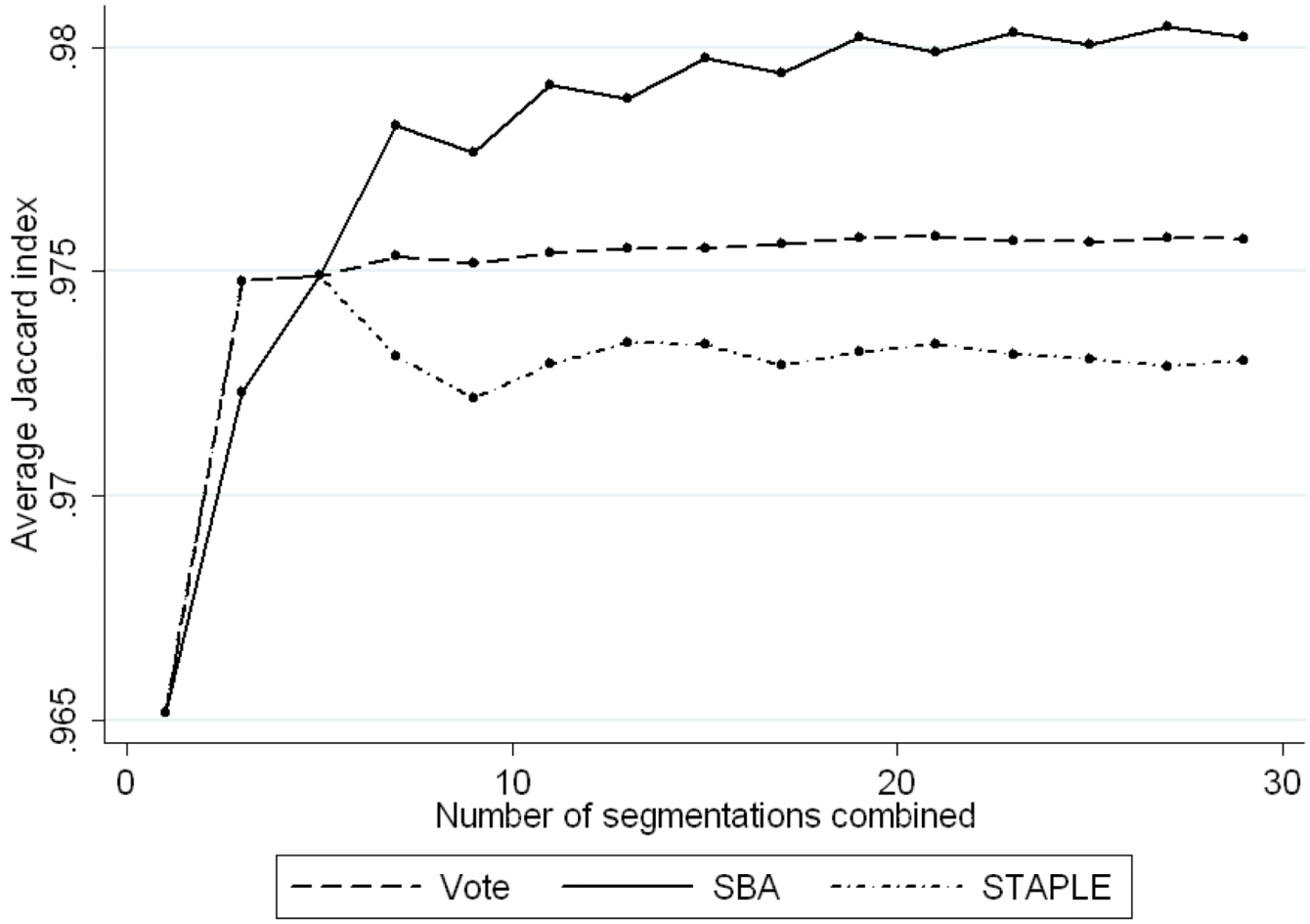
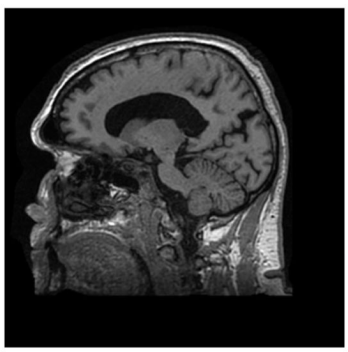
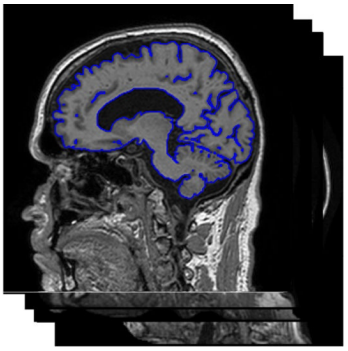


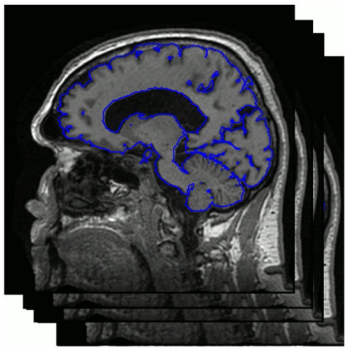
Figure 2.
MAPS parameter section: the figure shows the average Jaccard index of 'undilated MAPS-brains' using different numbers of best-matched atlases and label fusion techniques in a subset of 10 images.



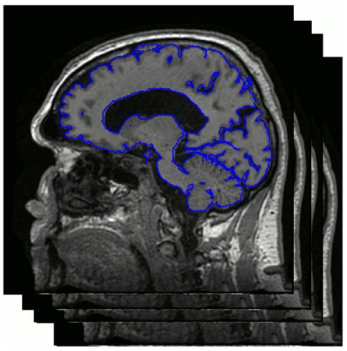
(a) Target image.



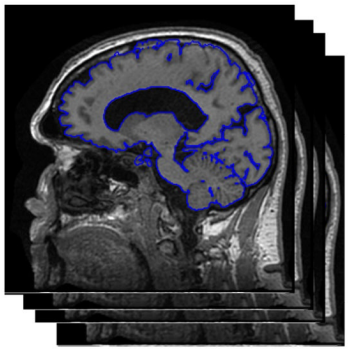
(b) Step 1. Template selection: selection of the 19 best-matched atlases from the template library.



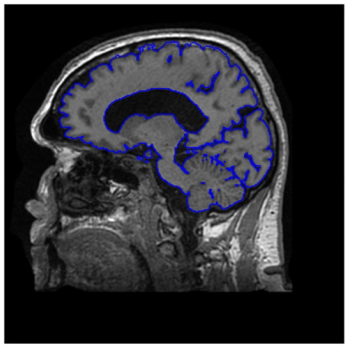
(c) Step 2.a. Label propagation: propagation of brain regions from the 19 best-matched atlases to the target image.



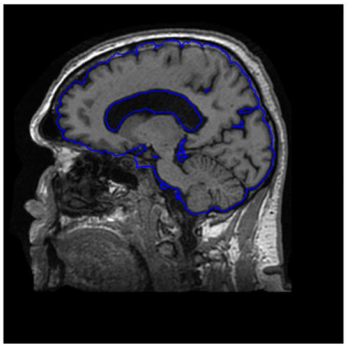
(d) Step 2.b. Label propagation: thresholding between 60% and 160% of the mean brain intensity.



(e) Step 2.c. Label propagation: 2-voxel conditional dilation between 60% and 160% of the mean brain intensity.



(f) Step 3.a. Label fusion: label fusion of the 19 different segmentations (‘undilated MAPS-brain’).



(g) Step 3.b. Label fusion: 2-voxel unconditional dilation (‘MAPS-brain’).

Figure 3.

Visual demonstration of MAPS. The sub-figures show the intermediate results of MAPS as described in Section 2.4.1 and Fig. 1.

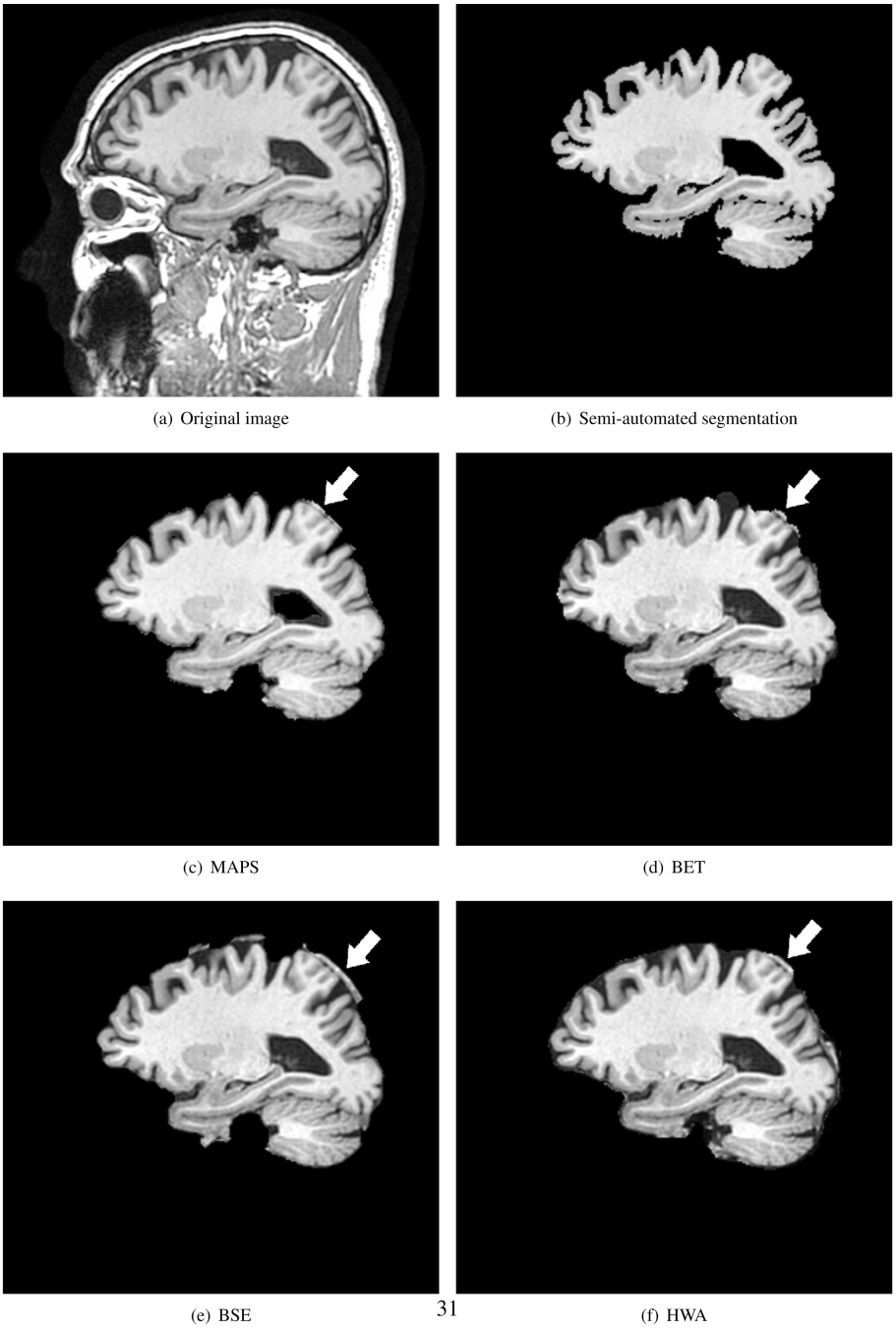


Figure 4.
Examples of whole brain extraction results of MAPS, BET, BSE and HWA of a 1.5T scan (ADNI subject ID: 126 S 0680). While all techniques had some errors in including non-brain (e.g. dura) voxels in some areas – the amount varied between methods (arrows).

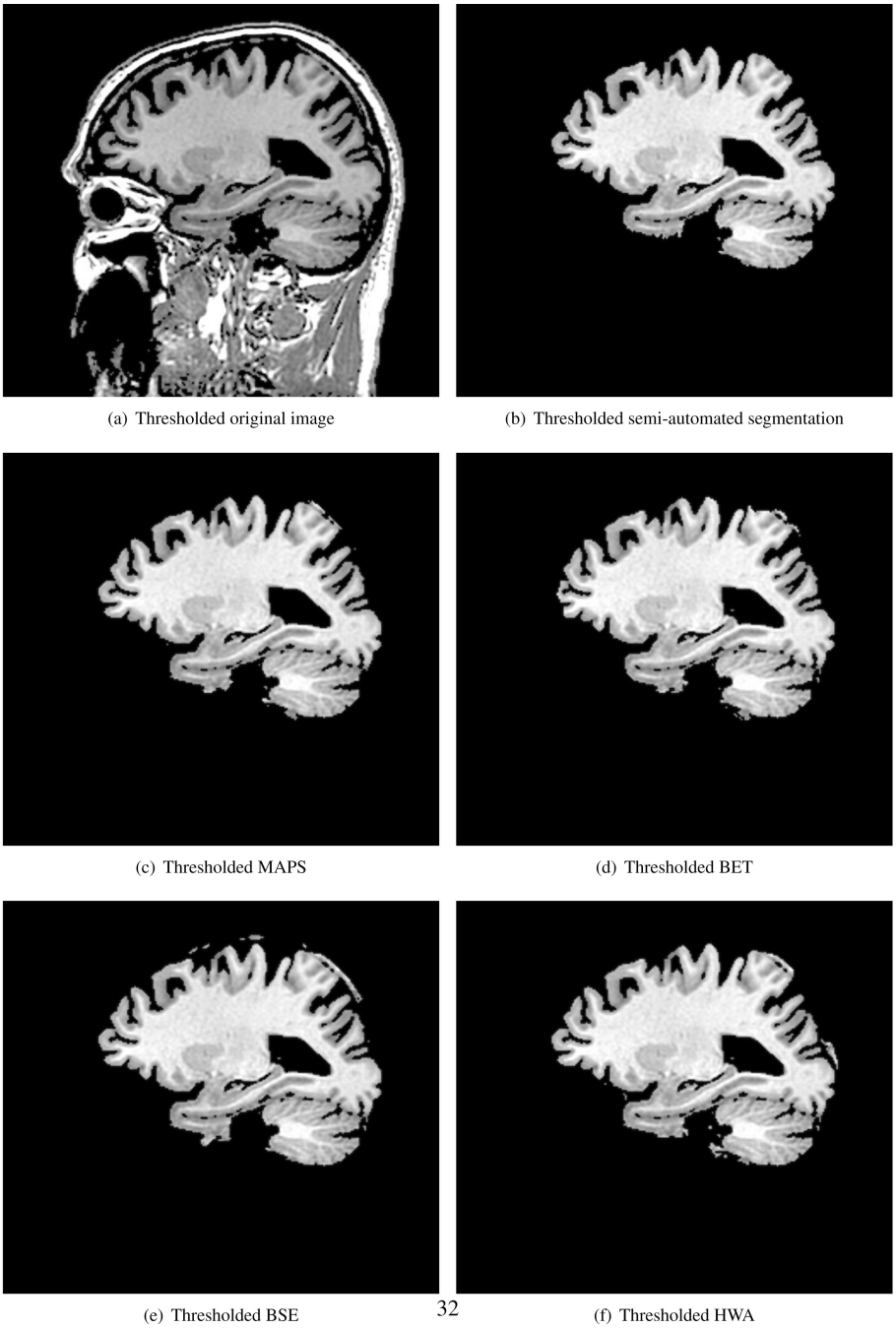


Figure 5.
Examples of whole brain extraction results of MAPS, BET, BSE and HWA of a 1.5T scan after thresholding using 60% of the mean intensity of the semi-automated whole brain segmentation (ADNI subject ID: 126 S 0680).

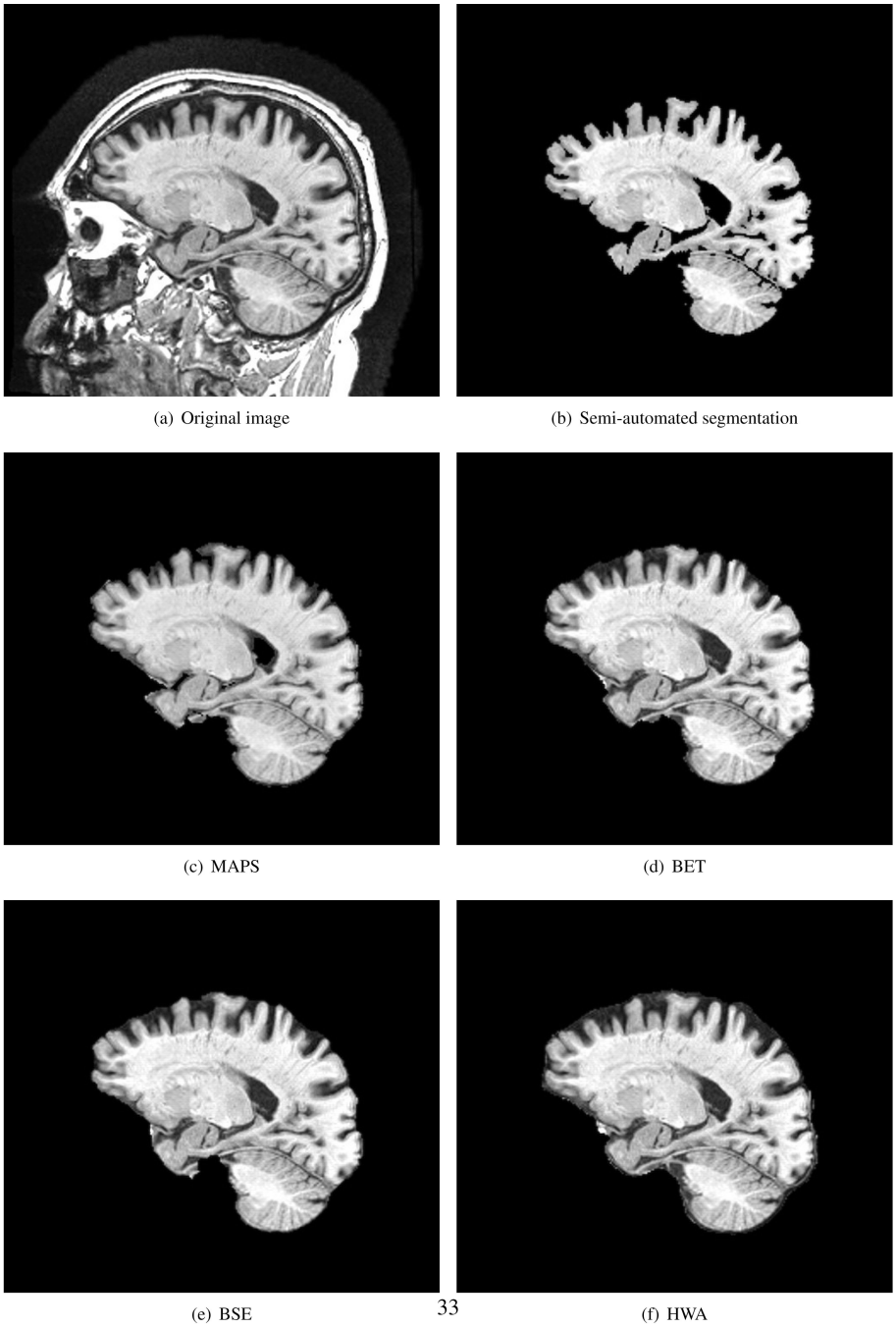


Figure 6.
Examples of whole brain extraction results of MAPS, BET, BSE and HWA of a 3T scan (ADNI subject ID: 037 S 1225).

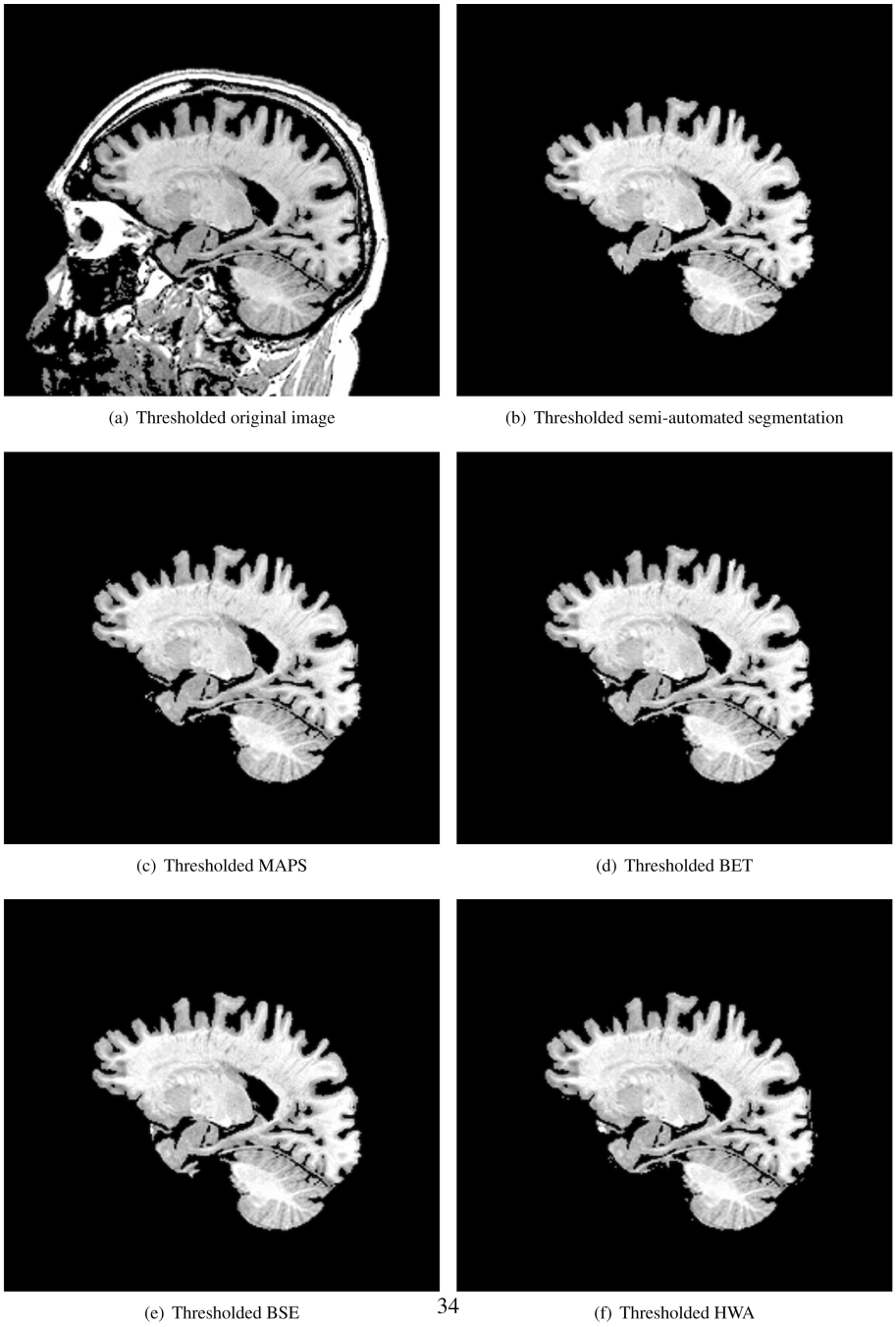


Figure 7.
Examples of whole brain extraction results of MAPS, BET, BSE and HWA of a 3T scan after thresholding using 60% of the mean intensity of the semi-automated whole brain segmentation (ADNI subject ID: 037 S 1225).

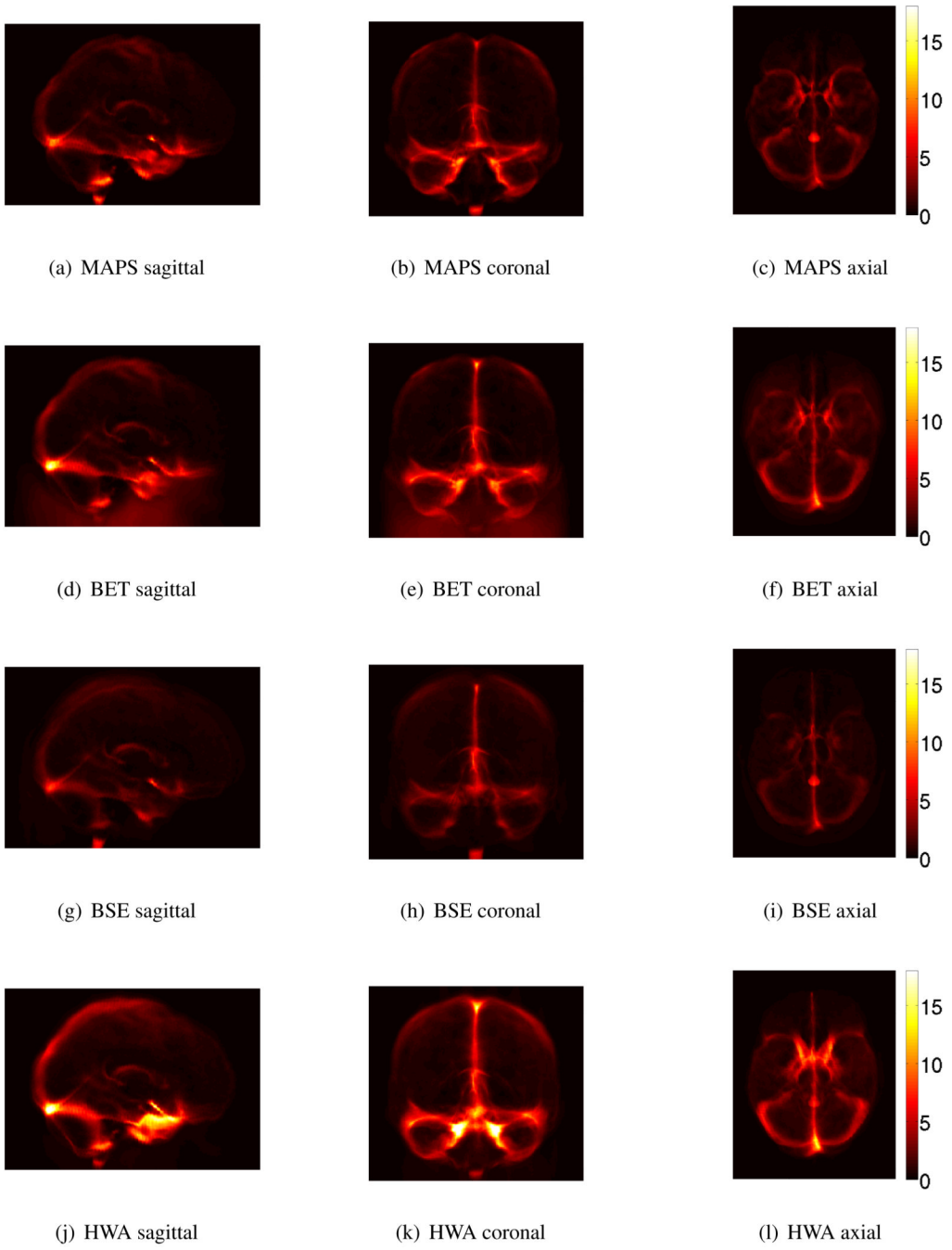


Figure 8.
Mean false positive maps of MAPS, BET, BSE and HWA from the segmentations of our whole dataset (682 1.5T and 157 3T scans). The colour maps show the average number of false positive counts (represented by the scales) in each projection plane.

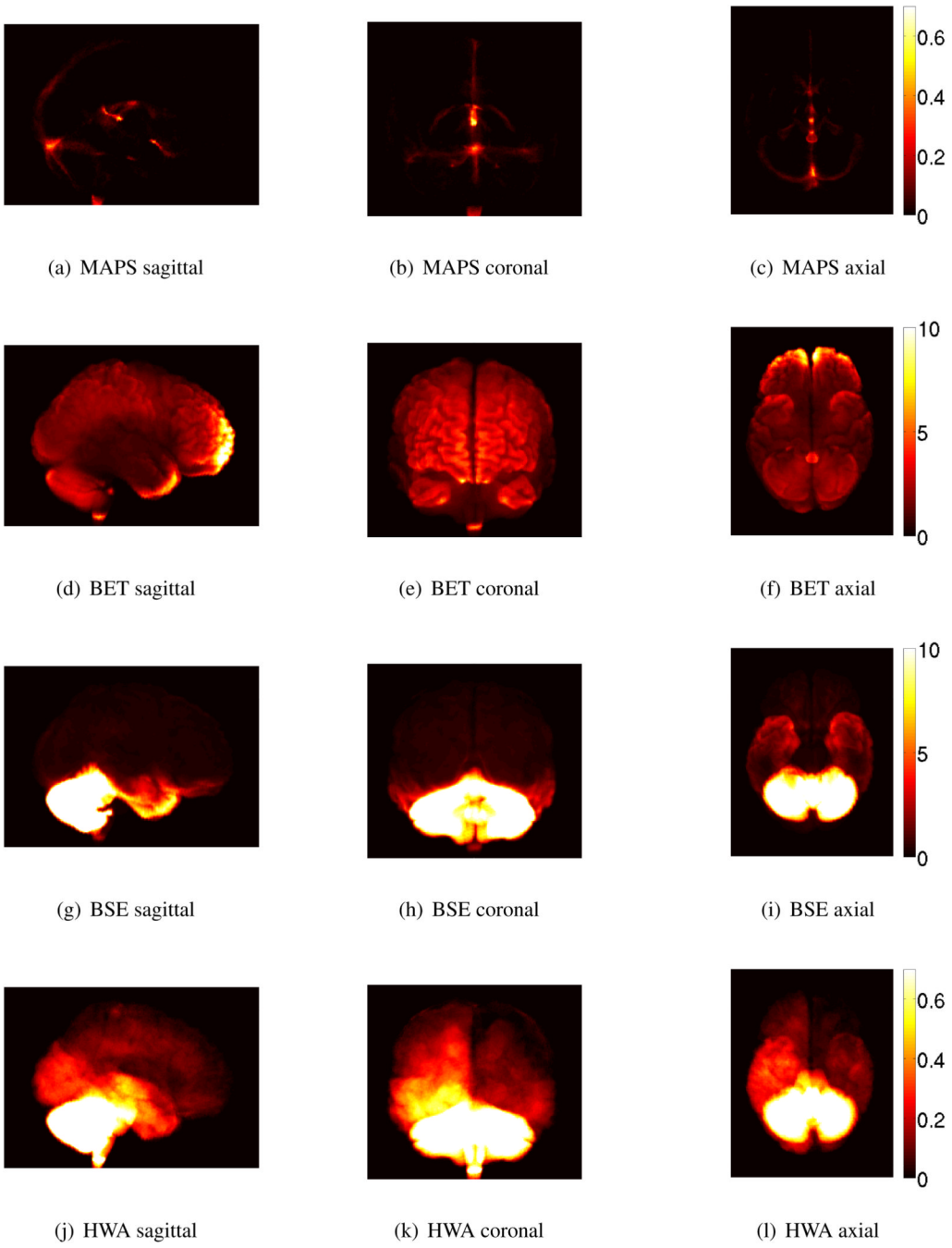


Figure 9.

Mean false negative maps of MAPS, BET, BSE and HWA from the segmentations of our whole dataset (682 1.5T and 157 3T scans). The colour maps show the average number of false negative counts (represented by the scales) in each projection plane. Note the differences in scale bar when comparing across these techniques; the scale bar for MAPS and HWA extend only to 0.6 whereas BET and BSE extend to 10.

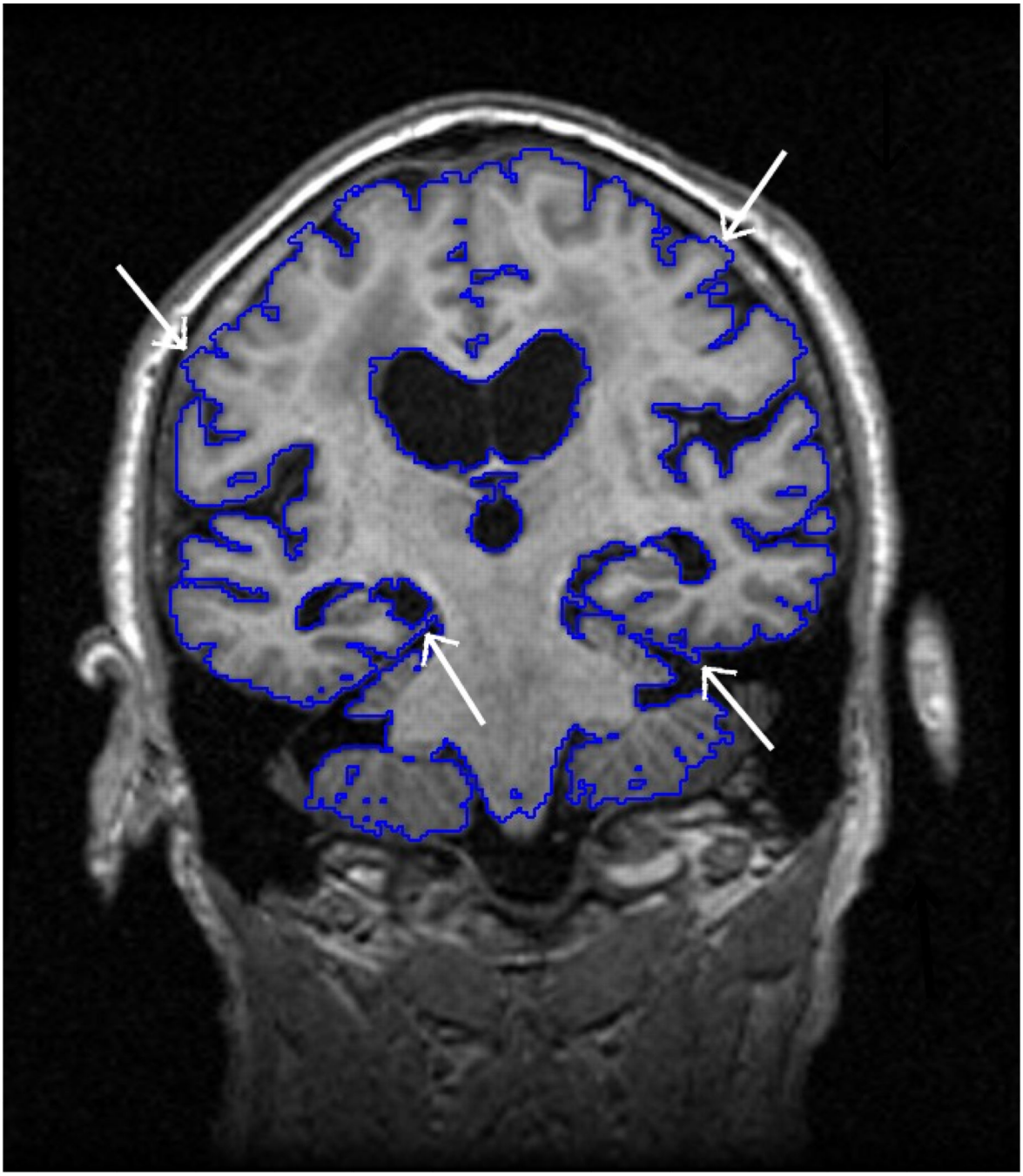


Figure 10.
Errors in a semi-automated segmentation. Extra dura and tentorial tissues were included in the segmentation (pointed by the white arrows).

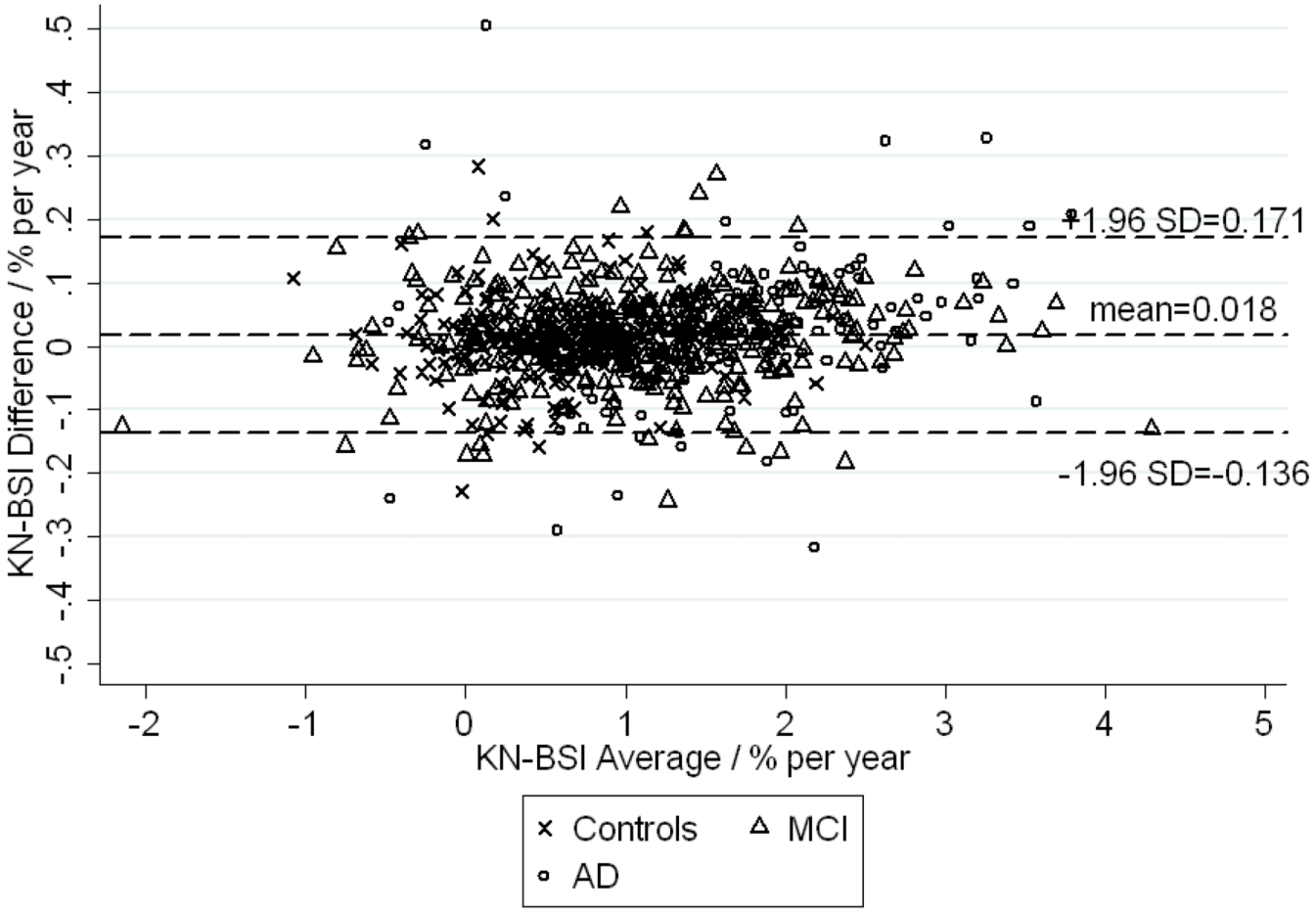


Figure 11.
Bland-Altman plot showing the agreement between brain atrophy measurement (as a percentage of the baseline brain volume) using KN-BSI calculated from semi-automated segmentations in baseline scans and propagated segmentations in 12-month follow-up scans (automated KN-BSI), and from 'undilated MAPS-brains' in baseline and 12-month follow-up scans (MAPS KN-BSI).

Table 1

A summary of automated brain extraction method comparison studies in chronological order from the literature.

Study	Sample size	Diagnostic group	Image acquisition
Shattuck et al. (2001)	20	Healthy subjects	T_1-weighted images from 1.5T scanner
Smith (2002)	45	Healthy subjects	35 T_1-, 6 T_2- and 4 Proton-density (PD)-weighted images from 1.5T and 3T scanners
Lee et al. (2003)	23	Healthy subjects	T_1-weighted images from 1.5T scanner
Boesen et al. (2004)	38	Healthy subjects	T_1-weighted images from 1.5T scanner
Ségonne et al. (2004)	43	Healthy subjects (14 young and 21 elderly) and subjects with dementia (2 AD and 6 with some form of dementia)	T_1-weighted images from 1.5T scanner
Fennema-Notestine et al. (2006)	32	Healthy subjects (8 young and 8 elderly), 8 unipolar depressed subjects and 8 AD subjects	T_1-weighted images from 1.5T scanner
Hartley et al. (2006)	296	Healthy subjects, 64 subjects with dementia and 59 subjects with infarcts	PD-weighted images from 1.5T scanner
Park and Lee (2009)	56	Healthy subjects	T_1-weighted images from 1.5T scanner
Shattuck et al. (2009)	40	Healthy subjects	T_1-weighted images from 1.5T scanner
Sadananthan et al. (2010)	68	Healthy subjects	T_1-weighted images from 1.5T and 3T scanners

Table 2

The demographics of the 682 subjects with 1.5T MRI scans and 157 subjects with 3T MRI scans.

	1.5T scans			3T scans		
	Control (N=200)	**MCI (N=338)**	**AD (N=144)**	**Control (N=53)**	**MCI (N=74)**	**AD (N=30)**
Mean (SD) age / years	76.0 (5.1)	74.9 (7.2)	75.4 (7.4)	75.3 (5.0)	74.9 (7.6)	74.8 (9.2)
Gender / male (%)	106 (53%)	214 (63%)	77 (53%)	19 (36%)	47 (64%)	11 (37%)

Table 3

The table shows the mean (SD) Jaccard index, false positive rate and false negative rate (5 controls, 5 MCI and 5 AD) between two different semi-automated brain segmentations by the same segmentor and by two different segmentors.

(a) Segmentations by the same segmentor	
	Jaccard index
Control	0.990 (0.005)
MCI	0.985 (0.005)
AD	0.991 (0.005)
All	0.988 (0.005)

(b) Segmentations by the two different segmentors	
	Jaccard index
Control	0.990 (0.004)
MCI	0.987 (0.002)
AD	0.990 (0.003)
All	0.989 (0.003)

Table 4

The mean (SD) Jaccard index of BET, BSE and HWA of the 18 randomly selected scans (one scan from each diagnostic group (Controls, MCI and AD) in each field strength (1.5T and 3T) from each scanner manufacturer (GE, Philips and Siemens)) from the parameter selection. The best parameters for each method are in bold. Note that only the top 5 BSE results are shown in the table.

Method	Parameters	Jaccard index
BET	default	0.634 (0.171)
	-R -f 0.5	0.719 (0.328)
	-S -f 0.5	0.643 (0.182)
	-B -f 0.5	0.887 (0.224)
	-B -f 0.4	0.910 (0.228)
	-B -f 0.3	**0.927 (0.187)**
	-B -f 0.2	0.921 (0.187)
	-B -f 0.1	0.881 (0.180)
	-B -f 0.0	0.761 (0.155)
BSE	**-n4 -d 20 -s 0.70 -p**	**0.917 (0.052)**
	-n 4 -d 19 -s 0.70 -p	0.914 (0.054)
	-n 10 -d 20 -s 0.70 -p	0.910 (0.148)
	-n5 -d 22 -s 0.70 -p	0.908 (0.139)
	-n 10 -d 21 -s 0.70 -p	0.908 (0.154)
HWA	default	0.961 (0.018)
	-less	**0.962 (0.018)**
	-more	0.960 (0.018)
	-less -atlas	0.932 (0.024)
	-more -atlas	0.228 (0.146)

Table 5

Median (1st–99th centile range) Jaccard indices, false positive rates and false negative rates of the automated whole brain segmentations of MAPS, BET, BSE and HWA using 1.5T scans of 200 controls, 338 MCI and 144 AD.

		Jaccard index (using thresholded segmentations)	False positive rate / % (using thresholded segmentations)	False negative rate / %
MAPS	Control	0.981 (0.041)	0.196 (0.440)	0.015 (0.226)
	MCI	0.981 (0.049)	0.177 (0.523)	0.011 (0.229)
	AD	0.980 (0.059)	0.192 (0.661)	0.007 (0.346)
	All	0.981 (0.049)	0.184 (0.509)	0.010 (0.242)
BET	Control	0.972 (0.909)	0.214 (11.2)	0.616 (82.9)
	MCI	0.969 (0.686)	0.193 (9.75)	0.967 (35.8)
	AD	0.965 (0.796)	0.201 (9.74)	0.903 (60.1)
	All	0.969 (0.826)	0.200 (10.3)	0.802 (60.3)
BSE	Control	0.954 (0.989)	0.116 (7.91)	2.03 (99.1)
	MCI	0.952 (0.172)	0.108 (0.945)	2.37 (16.2)
	AD	0.946 (0.270)	0.126 (2.42)	1.56 (12.5)
	All	0.953 (0.217)	0.116 (1.91)	2.17 (15.7)
HWA	Control	0.970 (0.143)	0.308 (0.676)	0.010 (11.1)
	MCI	0.971 (0.120)	0.289 (0.904)	0.009 (9.38)
	AD	0.968 (0.286)	0.293 (4.39)	0.007 (10.2)
	All	0.970 (0.126)	0.297 (0.894)	0.009 (7.22)

Table 6

Median (1st–99th centile range) Jaccard indices, false positive rates and false negative rates of the automated whole brain segmentations of MAPS, BET, BSE and HWA using 3T scans of 53 controls, 74 MCI and 30 AD.

		Jaccard index (using thresholded segmentations)	False positive rate / % (using thresholded segmentations)	False negative rate / %
MAPS	Control	0.980 (0.035)	0.173 (0.304)	0.015 (0.262)
	MCI	0.978 (0.048)	0.199 (0.514)	0.023 (0.213)
	AD	0.983 (0.040)	0.136 (0.444)	0.033 (1.13)
	All	0.980 (0.047)	0.177 (0.504)	0.019 (0.683)
BET	Control	0.969 (0.745)	0.168 (4.74)	1.05 (61.7)
	MCI	0.962 (0.721)	0.177 (6.68)	1.49 (44.6)
	AD	0.959 (0.137)	0.117 (0.353)	2.24 (14.1)
	All	0.965 (0.731)	0.161 (6.26)	1.30 (51.8)
BSE	Control	0.897 (0.977)	0.064 (0.376)	9.37 (99.2)
	MCI	0.899 (0.143)	0.089 (0.447)	9.18 (15.8)
	AD	0.905 (0.166)	0.057 (0.215)	8.78 (18.5)
	All	0.900 (0.550)	0.074 (0.420)	9.20 (56.1)
HWA	Control	0.965 (0.592)	0.295 (5.57)	0.007 (34.1)
	MCI	0.960 (0.849)	0.367 (9.68)	0.010 (49.2)
	AD	0.965 (0.581)	0.264 (9.75)	0.015 (43.7)
	All	0.962 (0.701)	0.321 (9.71)	0.010 (46.1)

Table 7

The comparison of the accuracy of MAPS, BET, BSE and HWA. The table shows the differences in the median (95% CI) of Jaccard index, false positive rate and false negative rate between the four automated brain extraction methods.

	Jaccard index (using thresholded segmentations)	False positive rate / % (using thresholded segmentations)	False negative rate / %
1.5T			
MAPS vs BET	0.012 * (0.011, 0.013)	−0.016 * (−0.022, −0.009)	−0.792 * (−0.876, −0.724)
MAPS vs BSE	0.028 * (0.021, 0.038)	0.068 * (0.058, 0.078)	−2.16 * (−3.09, −1.57)
MAPS vs HWA	0.011 * (0.009, 0.012)	−0.113 * (−0.122, −0.102)	0.002 (−0.001, 0.004)
HWA vs BET	0.001 (−0.000, 0.003)	0.097 * (0.086, 0.105)	−0.793 * (−0.878, −0.726)
HWA vs BSE	0.018 * (0.010, 0.028)	0.181 * (0.169, 0.192)	−2.16 * (−3.09, −1.57)
BET vs BSE	0.016 * (0.009, 0.026)	0.084 * (0.075, 0.095)	−1.37 * (−2.34, −0.807)
3T			
MAPS vs BET	0.015 * (0.012, 0.018)	0.015 * (0.000, 0.030)	−1.28 * (−1.52, −1.17)
MAPS vs BSE	0.079 * (0.072, 0.086)	0.102 * (0.086, 0.117)	−9.18 * (−10.0, −8.64)
MAPS vs HWA	0.018 * (0.015, 0.021)	−0.144 * (−0.184, −0.114)	0.008 * (0.003, 0.015)
HWA vs BET	−0.003 (−0.007, 0.001)	0.159 * (0.131, 0.199)	−1.29 * (−1.53, −1.18)
HWA vs BSE	0.062 * (0.055, 0.068)	0.246 * (0.220, 0.285)	−9.19 * (−10.0, −8.65)
BET vs BSE	0.065 * (0.058, 0.072)	0.087 * (0.072, 0.106)	−7.90 * (−8.77, −7.29)

* denotes statistical significance at $p < 0.05$.

Table 8

The comparison of the variability in accuracy of MAPS, BET, BSE and HWA. The table shows the differences in the 1st to 99th centile range (95% CI) of Jaccard index, false positive rate and false negative rate between the four automated brain extraction methods.

	Jaccard index (using thresholded segmentations)	False positive rate / % (using thresholded segmentations)	False negative rate / %
1.5T			
MAPS vs BET	−0.788 * (−0.891, −0.600)	−9.77 * (−10.4, −8.50)	−60.1 * (−88.5, −32.0)
MAPS vs BSE	−0.169 * (−0.581, −0.111)	−1.40 * (−3.47, −0.583)	−15.4 * (−34.5, −12.8)
MAPS vs HWA	−0.078 * (−0.139, −0.035)	−0.385 * (−6.72, −0.255)	−6.97 * (−12.4, −4.08)
HWA vs BET	−0.700 * (−0.847, −0.523)	−9.39 * (−10.1, −8.04)	−53.1 * (−84.8, −24.1)
HWA vs BSE	−0.091 * (−0.226, −0.010)	−1.02 * (−3.10, −0.174)	−8.45 * (−23.5, −1.61)
BET vs BSE	0.609 * (0.388, 0.771)	8.37 * (6.19, 9.40)	44.7 * (16.6, 75.3)
3T			
MAPS vs BET	−0.684 * (−0.708, −0.421)	−5.76 * (−6.31, −4.23)	−51.2 * (−61.5, −31.5)
MAPS vs BSE	−0.503 * (−0.950, −0.130)	0.084 * (0.037, 0.206)	−45.4 * (−49.0, −33.1)
MAPS vs HWA	−0.654 * (−0.813, −0.483)	−9.20 * (−9.36, −4.75)	−45.4 * (−49.0, −33.1)
HWA vs BET	−0.031 (−0.264, 0.478)	3.44 (−0.995, 9.29)	−5.78 (−28.2, 26.1)
HWA vs BSE	0.151 (−0.604, 0.612)	9.29 * (4.97, 9.53)	−10.0 (−83.0, 28.2)
BET vs BSE	0.182 (−0.808, 0.563)	5.84 * (4.36, 6.49)	−4.25 (−88.9, 37.5)

* denotes statistical significance at $p < 0.05$.

Table 9

Direct comparison of the ‘undilated MAPS-brains’ with semi-automated whole brain segmentations using 1.5T and 3T scans. The tables show the median (1st–99th centile range) Jaccard indices, false positive rates and false negative rates of the ‘undilated MAPS-brains’.

(a) 1.5T scans of 200 controls, 338 MCI and 144 AD			
	Jaccard index	**False positive rate /%**	**False negative rate / %**
Control	0.981 (0.047)	0.137 (0.395)	0.225 (3.68)
MCI	0.980 (0.062)	0.152 (0.492)	0.223(6.27)
AD	0.978 (0.061)	0.177 (0.492)	0.198 (6.27)
All	0.980 (0.053)	0.153 (0.457)	0.211 (4.76)

(b) 3T scans of 53 controls, 74 MCI and 30 AD			
	Jaccard index	False positive rate /%	False negative rate / %
Control	0.977 (0.058)	0.127 (0.261)	0.424 (6.12)
MCI	0.974 (0.083)	0.158 (0.453)	0.418 (8.41)
AD	0.971 (0.127)	0.123 (0.425	0.447 (13.8)
All	0.974 (0.106)	0.135 (0.462)	0.438 (11.2)

Table 10

Mean (SD) annualised brain atrophy measurement as a percentage of the baseline brain volume using KN-BSI calculated from semi-automated segmentations in baseline scans and propagated segmentations in 12-month follow-up scans (automated KN-BSI), and from ‘undilated MAPS-brains’ in baseline and 12-month follow-up scans (MAPS KN-BSI).

	Semi-automated KN-BSI	**MAPS KN-BSI**	**Difference (Semi-automated KN-BSI - MAPS KN-BSI) (95% CI), *p*-value**
Control (N=200)	0.608 (0.587)	0.596 (0.585)	0.012 (0.003, 0.021), $p = 0.008$
MCI (N=338)	1.128 (0.857)	1.110 (0.850)	0.017 (0.010, 0.0251), $p < 0.001$
AD (N=144)	1.566 (0.854)	1.541 (0.828)	0.025 (0.009, 0.043), $p = 0.005$