TECHNICAL REPORT

Evaluation of Automated Brain MR Image Segmentation and Volumetry Methods

Frederick Klauschen,[1]* Aaron Goldman,[2] Vincent Barra,[3] Andreas Meyer-Lindenberg,[2,4] and Arvid Lundervold[1]

[1]*Department of Biomedicine, Neuroinformatics and Image Analysis Laboratory, University of Bergen, Bergen, Norway*
[2]*National Institutes of Health, National Institute of Mental Health, Neuroimaging Core Facility, Genes, Cognition and Psychosis Program, Bethesda, Maryland*
[3]*LIMOS, UMR CNRS 6158, Blaise Pascal University Campus des Cezeaux, Aubiere, France*
[4]*Zentralinstitut f. Seelische Gesundheit, J5, Mannheim, Germany*

Abstract: We compare three widely used brain volumetry methods available in the software packages FSL, SPM5, and FreeSurfer and evaluate their performance using simulated and real MR brain data sets. We analyze the accuracy of gray and white matter volume measurements and their robustness against changes of image quality using the BrainWeb MRI database. These images are based on "gold-standard" reference brain templates. This allows us to assess between- (same data set, different method) and also within-segmenter (same method, variation of image quality) comparability, for both of which we find pronounced variations in segmentation results for gray and white matter volumes. The calculated volumes deviate up to >10% from the reference values for gray and white matter depending on method and image quality. Sensitivity is best for SPM5, volumetric accuracy for gray and white matter was similar in SPM5 and FSL and better than in FreeSurfer. FSL showed the highest stability for white (<5%), FreeSurfer (6.2%) for gray matter for constant image quality BrainWeb data. Between-segmenter comparisons show discrepancies of up to >20% for the simulated data and 24% on average for the real data sets, whereas within-method performance analysis uncovered volume differences of up to >15%. Since the discrepancies between results reach the same order of magnitude as volume changes observed in disease, these effects limit the usability of the segmentation methods for following volume changes in individual patients over time and should be taken into account during the planning and analysis of brain volume studies. *Hum Brain Mapp 30:1310–1327, 2009.* © 2008 Wiley-Liss, Inc.

Key words: structural MRI; segmentation; accuracy; volumetry

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsors: NIMH/IRP and The Research Council of Norway (Aur06-11), France-Norway AURORA Project.

*Correspondence to: Frederick Klauschen, National Institutes of Health, 9000 Rockville Pike, 10-11N311, Bethesda, MD 20892-1892, USA. E-mail: fklauschen@mail.nih.gov

Received for publication 1 June 2007; Revised 28 March 2008; Accepted 31 March 2008

DOI: 10.1002/hbm.20599

Published online 6 June 2008 in Wiley InterScience (www.interscience.wiley.com).

INTRODUCTION

The ability to segment and quantify brain tissues and anatomical structures from 3D MRI acquisitions has increasing importance in the study of brain development [e.g., Counsell and Boardman, 2005; Nishida et al., 2006], neurodegeneration [e.g., Cordato et al., 2005; Sepulcre et al., 2006], and dementia [e.g., Fotenos et al., 2005; Ridha et al., 2006], and in the assessment of neurological [e.g., Meyer-Lindenberg et al., 2005; Ciumas and Savic, 2006] and psychiatric [e.g. Henriksson et al., 2006; Honea et al., 2005; Marcelis et al., 2006; Sporn et al., 2003] disorders. Brain morphometry from MRI is also used to investigate brain-behavioral relationships such as IQ and memory performance [e.g. Fjell et al., 2005; Frangou et al., 2004; Walhovd et al., 2005], as well as genetic influences [e.g., Meyer-Lindenberg et al., 2006; Pezawas et al., 2005].

The history of tissue segmentation from brain MR images started with the seminal work of Vannier [1985] by adopting statistical classification software from NASA. During the last 20 years, there has been an enormous methodological development in the field of brain segmentation as well as image acquisition techniques (for reviews see: Zijdenbos and Dawant [1994]; Saeed [1998]; Pham et al. [2000]; Duncan et al. [2004]).

The simplest but most time-consuming method is slice-by-slice manual tracing. This operator-dependent method is still being used, mainly as a "gold-standard" reference method for whole brain and GM/WM/CSF segmentations (cf. the Internet Brain Segmentation Repository [http://www.cma.mgh.harvard.edu/ibsr]), for segmentation of subcortical structures [e.g., Szabo et al., 2006] such as the amygdala and hippocampi [e.g., Barnes et al., 2006], and for correction of local misclassifications as part of the processing chain in semiautomated segmentation methods.

Semiautomated unsupervised and supervised pattern recognition techniques using contextual classification methods and estimation of Mahalanobis distances between tissue types in feature space was proposed by Taxt and Lundervold [1994] for segmentation of multispectral MRI from the brain. A similar multispectral automated discriminant analysis approach, including fuzzy classification, was taken by Harris et al. [1999] and Amato et al. [2003], introduced independent component analysis (ICA) and tissue-specific nonparametric probability density functions into multispectral brain image segmentation. However, with the introduction of 3D MRI acquisitions on modern scanners of today, enabling whole brain coverage, high spatial resolution, and good contrast-to-noise ratios within a few minutes measurement time [e.g., Magnotta et al., 2006], the majority of brain segmentation and morphometric studies make use of 3D image registration and electronic brain atlases (templates) with prior tissue probabilities for their voxel classification, rather than multiple pulse sequences and multispectral analysis. Several software packages for brain segmentation currently in use (e.g., SPM2/SPM5, FSL, FreeSurfer, and BrainVoyager) employ such a priori information.

Recent developments also allow integration of volume- and surface-based methods [Kim et al., 2005; Makris et al., 2006] to perform cortical topographic measurements, and mathematics-oriented investigators have started to use level-sets, PDEs, and variational methods for MR image processing and brain segmentation [e.g., Cates et al., 2004; Droske et al., 2005; Leow et al., 2005; Lie et al., 2006].

Apart from these theoretical developments, it is also of great practical importance to investigate the performance of competing methods. Several studies have been conducted where different skull-stripping and brain segmentation algorithms are compared, both with each other, and with manual tracing, or realistic digital brain phantoms [Barra and Boire, 2001; Byrum et al., 1996; Cuadra et al., 2005; Fennema-Notestine et al., 2006; Good et al., 2002; Grabowski et al., 2000; Greenspan et al., 2006; Heckemann et al., 2006; John et al., 2003; Kovacevic et al., 2002; Lemieux et al., 2003; Moretti et al., 2000; Rehm et al., 2004; Toga and Thompson, 2003; Wang and Doddrell, 2002; Warfield et al., 2004; Zaidi et al., 2006; Bezdek et al., 1993]. Also, the impact of MR image acquisitions protocols on tissue segmentation results and brain volumes has been investigated [e.g., Lundervold et al., 2000; Clark et al., 2006], and one study specifically addressed reproducibility of volumetry results over time of Chard et al. [2002].

In the search for biological causes of brain volume differences between diagnostic groups or individual changes during time in longitudinal studies, the variations due to MRI measurement technique, data quality, and image segmentation procedure should be explored and accommodated in the data analysis. In the recent study by Clark et al. [2006], the choice of segmentation algorithm had the largest impact on variability, whereas the choice of a pulse sequence had the second largest impact. Moreover, the classification of gray matter was the most variable, and the optimal protocol could differ across tissue types. In a systematic review and meta-analysis of 66 papers comparing brain volume in patients with a first psychotic episode with volume in healthy controls, Grant Steen et al. [2006] concluded that a major problem seems to be that the volumetric loss in patients, which is no more than 4% per year, may be close to the limit of detection by MRI. Thus, poor precision or low accuracy in even a subset of volumetric studies would lead to a lack of consensus among the various studies.

In this work, we evaluated the performance of three widely used software packages for brain volumetry: SPM5 [Ashburner and Friston, 2005], FreeSurfer [Dale et al., 1999; Fischl et al., 1999], and FSL [Smith, 2004]. The aim of our work was on one side to explore volumetric variation caused by algorithmic effects and moreover, to address the question to what extent variations of image quality (noise and intensity inhomogeneities) influence volumetric results even when the same method is used. The evaluation of this within-segmenter performance was achieved by the use of synthetic data, namely the Montreal Neurological

Institute BrainWeb digital brain phantom that provides MRI data sets with varying image quality based on one gold-standard tissue segmentation mask. We complemented the inter-segmenter analysis by comparing the volumetry results obtained for multiple-anatomical-model BrainWeb data sets, real 3D MRI data recorded in a study of normal aging, and images from the OASIS database (http://www.oasis-brains.org).

MATERIALS AND METHODS

Statistical Parametric Mapping SPM5

The SPM software package SPM5 is a suite of MATLAB (The MathWorks, Natick, MA) functions and subroutines (including some C code) that implements statistical methods for analysis of functional and structural neuroimages. The segmentation process in SPM5 is an integrated generative modeling approach, in which tissue segmentation, intensity normalization, and nonlinear warping are performed within the same mixture of Gaussian model [Ashburner and Friston, 2005]. The segmentation process does not only work on single voxels but takes contextual signal intensity information into account that is encoded in template images containing prior probabilities for GM, WM, and CSF. These spatial priors are also deformed to the subject brain to allow registration to a standard space. For a successful spatial normalization, it is important that the tested brain is similar to the template brain, e.g., when examining children's brains a special template data set has to be provided. In this study, the standard template brain included in SPM5 was used. In SPM5, classification is probabilistic in the sense that a probability value of belonging to each of the classes is assigned to each voxel. These probability values sum to unity. Total tissue volumes were calculated by adding up, over all voxels, the assigned probability of the given class, and then by multiplying by the known voxel volume [according to Lüders et al., 2002].

Segmentations in this study were performed using SPM5 Revision 546, released on June 5, 2006. Because initial proximity to the template was observed to have a strong effect on result quality, a rigid-body rotation and translation was first performed on all subjects using the coregister function in SPM5, and the SPM5 T1 template as the reference image. All subjects were then segmented using the default templates (a modified version of the ICBM Tissue Probabilistic Atlas, located at http://www.loni.ucla.edu/ICBM/ICBM_Probabilistic.html) and parameters for this version. Specifically, these parameters included 2 Gaussians each for WM, GM, and CSF and 4 Gaussians for everything not fitting these categories, a warping regularization value of 1, a warp frequency cutoff of 25, very light regularization (0.0001), a 60-mm cutoff for the full width at half maximum (FWHM) of Gaussian smoothness of bias, and a sampling distance of 3. SPM5 segmentation results are output as probability maps with voxel values between 0 and 255. When generating binary images (for visualization of STAPLE/VOTING), voxels with a probability of ≥ 0.5 (i.e., 128) were counted as members of that particular class. Using a value of 0.5 for the class membership decision prevents voxels in the border region between white and gray matter to be classified as both gray and white matter.

SPM5 can be obtained from http://www.fil.ion.ucl.ac.uk/spm/.

FreeSurfer

FreeSurfer is a set of tools for automated surface reconstruction and analysis, which extracts white matter and pial surfaces, computes measures such as thickness and sulcal depth, and performs cross-subject analysis using spherical registration [Dale et al., 1999; Fischl et al., 1999]. FreeSurfer also parcellates the cortex into gross anatomical regions and produces statistics on thickness, area, and volume for each region [Fischl et al. 2004]. In addition to its surface reconstruction package, FreeSurfer includes a sophisticated automated segmentation algorithm, which delineates gross brain anatomy into a series of cortical and subcortical labels. Briefly, structures are labeled using a complex algorithm combining information on image intensity, probabilistic atlas location, and the local spatial relationships between subcortical structures [Fischl et al., 2002, 2004]. For this purpose, calculated volumes for these subcortical labels were summed to derive estimates of total gray and white matter volume. For gray matter, we calculated the total combined volume of cerebral and cerebellar cortex, hippocampus, amygdala, caudate, putamen, globus pallidus, nucleus accumbens, thalamus, and ventral diencephalon, and for white matter we summed cerebral and cerebellar white matter, brain stem, and white matter hypointensities. Additionally, we compared cortical gray matter estimates from the segmentation algorithm with totals derived from the cortical parcellation algorithm. For this purpose, the total volumes of each parcellation label except unknown and corpus callosum were added together to obtain one cortical gray matter value.

The reported comparisons were performed using the FreeSurfer Stable 3.0.2 release, using the default processing stream (recon-all -all). Data were visually inspected, and manual interventions were performed where automated steps had failed. These included manual alignment to the talairach template in cases where automated registration was poor, and adjustments to the watershed threshold to restore areas of the brain that were erroneously removed during skull stripping. The FreeSurfer software and its documentation can be downloaded from http://surfer.nmr.mgh.harvard.edu.

In this study, the volume-based segmentation approach of FreeSurfer is used. As an addition, we compared the surface- and volume-based approaches for cortical gray matter (see section "FSL").

FSL

FSL (http://www.fmrib.ox.ac.uk/fsl/) is a library of image analysis and statistical tools for fMRI, MRI, and DTI brain imaging data. It is composed of several modules including structural tools (BET, brain extraction; FAST, tissue segmentation; FLIRT, linear registration; FUGUE B0, unwarping; SIENA, brain change analysis), functional (FEAT, model-based FMRI analysis; MELODIC, probabilistic ICA temporal model-free FMRI analysis) and connectivity (FDT, diffusion and tractography, and TBSS, VBM-like analysis with FA data) components. The segmentation tool of FSL (FAST, FMRIB's Automated Segmentation Tool) segments a 3D image of the brain into different tissue types (gray matter, white matter, CSF, etc.), while also correcting for RF inhomogeneities. The whole process is fully automated and can also produce a bias field-corrected input image and a probabilistic and/or partial volume tissue segmentation, from which tissue volumes were computed. The FAST algorithm is based on a hidden Markov random field (MRF) model and an associated expectation-maximization algorithm [Zhang et al., 2001]. It can be processed in various ways: from scratch (without any a priori model, only using the MRF), using a priori information (a priori maps created from averaging many segmentations) for both initialization and posteriors for the algorithm, or allowing the estimation of partial volume compartments. For the segmentation results presented here, FAST (Version 3.53, part of FSL Version 3.3/4.0) was used with these different parameter settings, using probability maps (default settings), partial volume estimation (fast -e), and a priori information (fast -A). We denote these three possible uses PBMAP, PVE, and APRIORI, respectively. If not otherwise noted, PBMAP segmentation is used because this is the default setting in FAST.

We used both BET (bet2, Brain Extraction Tool Version, part of FSL) and BSE (brain surface extraction), which is part of BrainSuite (http://brainsuite.usc.edu) and recommended by Fennema-Notestine et al., 2006.

MRI Brain Data

BrainWeb—Simulated brain data

Simulated MRI data sets were used as test data, generated with the Internet connected MRI Simulator at the McConnell Brain Imaging Centre in Montreal http://www.bic.mni.mcgill.ca/brainweb/. The data sets are based on an anatomical model of a normal brain that results from registering and preprocessing of 27 scans from the same individual with subsequent semiautomated segmentation. In this data set, the different tissue types are well-defined, both "fuzzy" and "crisp" tissue membership are allocated to each voxel. From this tissue-labeled brain volume, the MR simulation algorithm, using discrete-event simulation of the pulse sequences based on the Bloch equations, predicts signal intensities and image contrast in a way that is equivalent to data acquired with a real MR-scanner (resembling ~1.5T images). Both sequence parameters and the effect of partial volume averaging, noise, and intensity nonuniformity are incorporated in the simulation results [Cocosco et al., 1997; Collins et al., 1998; Kwan et al., 1999].

Ten data sets (T1, voxel size: 1 mm^3) with variations of the parameters "noise (n)" (ranging from 1 to 9%) and "intensity nonuniformity (rf)" (20 and 40%) were chosen: n1rf20, n1rf40, n3rf20, n3rf40, n5rf20, n5rf40, n7rf20, n7rf40, n9rf20, n9rf40. This selection covers the whole range of the parameter values available in BrainWeb so that the comparability with real data can be considered as sufficient to test the robustness of the different methods at varying image qualities.

To obtain the "true", i.e., reference volumes, the voxels labeled as gray and white matter in the discrete brain phantom (noise = 0%, RF = 0%) were counted. The additional 20 simulated BrainWeb data sets that were used are each based on an anatomical model of an individual normal brain and thus allow us to test the segmenter performance when image quality is constant (3% noise, 0% intensity-inhomogeneity) and anatomy is varied. In the following, these data sets are referred to as "multiple-anatomical-model" data. For details see Aubert-Broche et al. [2006] and http://www.bic.mni.mcgill.ca/brainweb/anatomic_normal_20.html. The discrete model data sets available online have a higher resolution (362 × 434 × 362) than the simulated data sets (181 × 256 × 256). To be able to perform a voxel-wise comparison between the discrete and simulated data, discrete data sets resampled to the resolution of the simulated data were kindly provided upon our request by B. Aubert-Broche.

Real data

Nine data sets, selected from a sample of healthy volunteers (Female 53 yrs, Male 72 yrs, F 71 yrs, F 52 yrs, M 74 yrs, M 55 yrs, M 57 yrs, M 62 yrs, F 54 yrs) participating in a study of cognitive aging, brain function, and genetic markers, were recorded on a 1.5 T GE Signa Echospeed scanner with a standard 8-channel head coil, using 256 × 256 × 124 dual-volume SAG T1 3D FSPGR IR prepared acquisitions (TR/TE/TI/FA = 9.5/2.2/450/7 deg) at voxel-size 0.94 × 0.94 × 1.4 mm^3. All subjects gave their written informed consent to participate in the study, which was approved by the Regional Committee for Medical Research Ethics of Southern Norway.

The 48 MP-RAGE data sets were obtained from the Open Access Series of Imaging Studies (OASIS) at http://www.oasis-brains.org, disc1 (OASIS datasets number: 1 (74 yrs), 2 (55 yrs), 3 (73 yrs), 4 (28 yrs), 5 (18 yrs), 10 (74 yrs), 13 (81 yrs), 19 (89 yrs), 28 (86 yrs), 31 (88 yrs), 32 (89 yrs), 33 (51 yrs), 35 (27 yrs), 52 (78 yrs), 53 (83 yrs), 56 (72 yrs), 61.1 and 61.2 (20 yrs), 64 (77 yrs), 65 (90 yrs), 67 (71 yrs), 75 (83 yrs), 80.1 and 80.2 (25 yrs), 83 (90 yrs), 85 (70 yrs), 92.1 and 92.2 (22 yrs), 101.1 and 101.2 (29 yrs), 106 (81 yrs), 110 (84 yrs), 111.1 and 111.2 (23 yrs), 117.1 and 117.2 (25 yrs), 122 (83 yrs), 134 (80 yrs), 137 (87 yrs), 145.1 and

TABLE I. Discrete reference data BrainWeb-volumes in mm^3

Data set	gm	wm	gm + wm
BrainWeb 1–10	902910	674780	1577690
4	963649	646859	1610508
5	1011083	608520	1619603
6	939339	676213	1615552
18	1053434	572797	1626231
20	997071	604577	1601648
38	1019186	590372	1609558
41	1017517	605186	1622703
42	1031546	574882	1606428
43	1108560	662370	1770930
44	1009779	615385	1625164
45	956830	647479	1604309
46	975585	605093	1580678
47	982427	630089	1612516
48	893987	671902	1565889
49	924584	743026	1667610
50	910180	632713	1542893
51	965621	606954	1572575
52	979283	619042	1598325
53	981646	571190	1552836
54	986608	575658	1562266

BrainWeb 1-10: variable quality data; Data sets 4 to 54: multiple-anatomical-model data.

145.2 (34 yrs), 150.1 and 150.2 (20 yrs), 156.1 and 156.2 (20 yrs), 185 (78 yrs), 191.1 and 191.2 (21 yrs). We use the abbreviations O1, O2, etc. for the OASIS data sets.

Miscellaneous

The software authors were contacted to verify that the approach we use in this publication is recommended.

The postprocessing of the segmentation results (confusion matrix analysis, STAPLE/VOTING, visualization, etc.) was implemented and performed using the C++ programming language, MATLAB (http://www.mathworks.com), and NIFTI-Tools (http://www.mathworks.com/matlabcentral/fileexchange).

RESULTS

BrainWeb

Table I contains the values for the volumes of the (unsimulated) reference brains.

Confusion Matrix Analysis

To assess the sensitivity and specificity of the methods, which are not necessarily reflected by the simple volume counts, we calculated the confusion matrices. In the optimal case of perfect classification, the confusion matrix would be the identity matrix. The confusion matrix is read as follows: When i describes the row and j the column, then the confusion matrix element c_{ij} contains the relative number of voxels that belong to class i and are classified as j, i.e., for a given i and any $j \neq i$, c_{ii} contains the relative number of true positive, c_{jj} the true negative, c_{ij} the false negative, and c_{ji} the relative number of false positive voxels. From these values, the sensitivity [= tp/(tp + fn)][1] and specificity [= tn/(tn + fp)] of the method can be computed (see later). Figure 1 shows a multiple confusion plot for the BrainWeb data of variable quality, providing the results for all three segmenters for each data set. Table II shows the average confusion matrices. Overall, SPM5 shows the highest similarity with the identity matrix with a mean probability of gray matter to be classified as gray matter of 89.1%, which is 6.3%/8.8% more sensitive than for FSL/FreeSurfer. The highest off-diagonal values can be found for all segmenters for the gm-wm, wm-gm, and gm-om[2] fields with FreeSurfer showing the highest probability of classifying gray as white matter (μ = 12.9%) and FSL being more likely to label gray as non-white non-gray tissue (μ = 10.4%).

Figure 2 shows the multiple confusion matrix plots for the 20 multiple-anatomical model BrainWeb data sets. Here, SPM5 and FSL results are similar and relatively close to the identity matrix when compared with FreeSurfer (the average probability of gray matter to be classified as gray matter for SPM5 and FSL is 91.3% and 90.4%, but only 83.2% for FreeSurfer). The main difference between the two groups is that the FSL matrix gm-gm field is closer to the identity matrix in the multiple-anatomical-model data than in the variable-quality group. The confusion matrix analysis suggests a dependency of the SPM5 results on data quality (decreasing quality from data set 1 to 10) for gm-gm (Spearman rank correlation $r_s = -0.85$) and of both SPM5 and FSL results for gm-wm ($r_s^{\mathrm{SPM5}} = 0.95$ and $r_s^{\mathrm{FSL}} = 0.90$). As expected no such observation can be made in the multiple-anatomical model group because the simulation parameters are constant, only the anatomical models used for each simulated data set are different. The overall configuration of the confusion matrices does not change between the two groups, i.e., the multiple-anatomical model data group also shows the highest off-diagonal values in the fields gm-wm, wm-gm, and gm-om, and both diagonals show a high similarity (see Table III).

Based on the BrainWeb data, we computed the average sensitivities and specificities for the three methods as shown in Table IV.

Figure 3 shows the segmentation results for an exemplary slice (subject 06, slice 73) overlaid with the reference data, and Supplementary Figure 2 shows a similar visualization for slice 100. This example supports the trend observed in the confusion matrices that SPM and FSL have a higher accuracy than FreeSurfer, for which the results show a larger misclassification with a slight anterior–posterior asymmetry. FreeSurfer underestimates gray (ratio of false negative voxels and reference number of voxels = 0.14) and overestimates white matter (ratio of false positive voxels and reference number of voxels = 0.1). This aspect of FreeSurfer segmentation is replicated in the volume count results for the whole brain that consistently show an overestimation of white and underestimation of gray matter.

[1] tp, tn, fp, fn: true/false positive/negative.

[2] om: other matter, i.e., all non-white and non-gray matter.

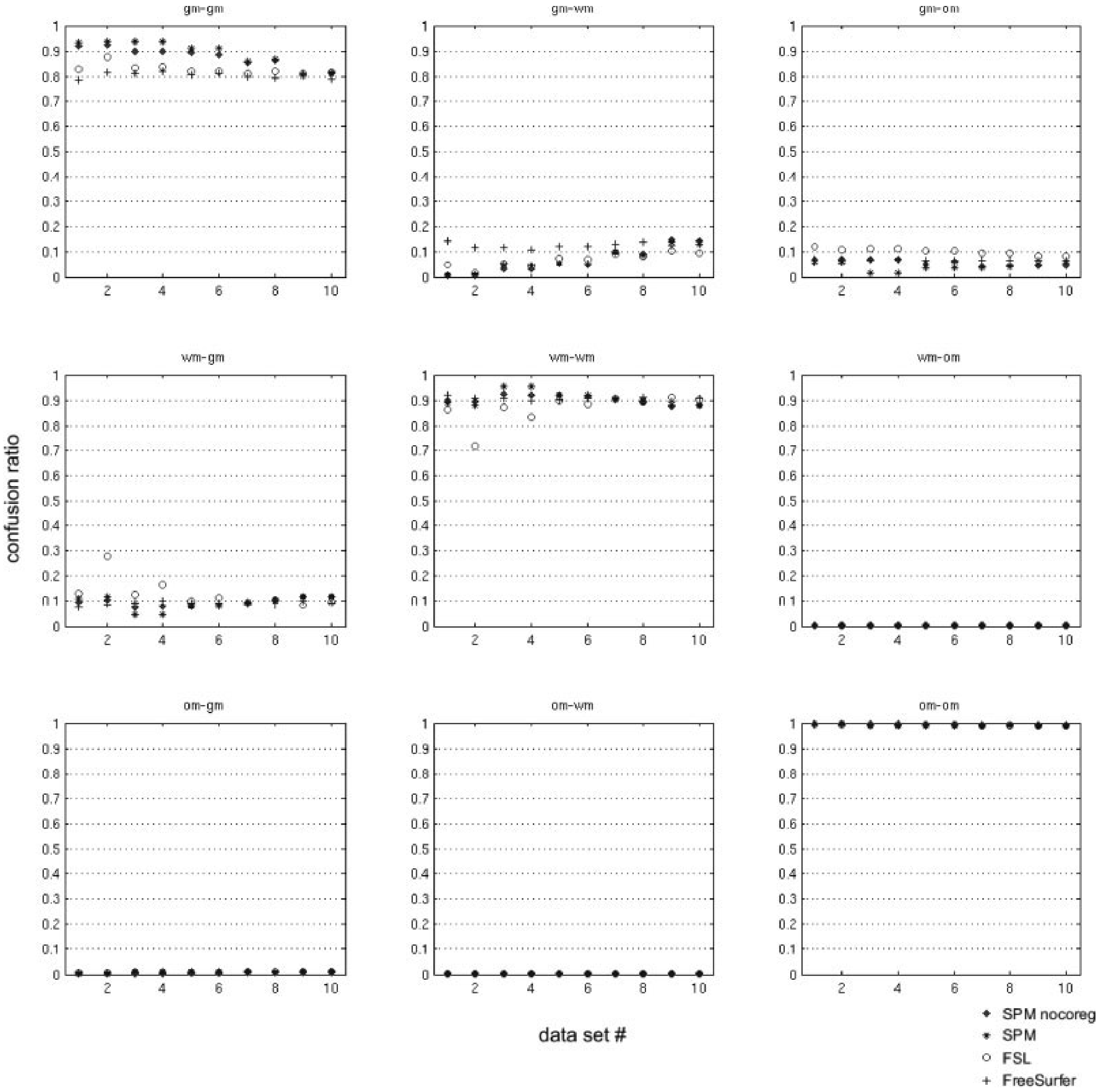

Figure 1.

Multi-confusion-matrix (variable-quality BrainWeb data): Each element of the confusion matrix contains the results for the three segmenters for all 10 data sets (Labels 1, 2, 3, etc. correspond to n1rf20, n1rf40, n3rf20, etc.). SPM and SPM _nocoreg indicate if data sets were or were not registration-transformed prior to segmentation; gm: gray; wm: white; om: other (non-gm, non-wm) matter; "im-jm": values denote the relative number of voxels belonging to class "i" classified as "j" (confusion ratio).

Optimization by data fusion

Supplementary Figure 1 and Supplementary Table II show the results of an (exemplary) optimization attempt based on fusing the three available segmentations. Two different fusing methods were used: (1) VOTING, i.e., voxels are assigned a specific class label if the majority of segmentations (at least two) agree on the label and (2) STAPLE ("Simultaneous Truth And Performance Level Estimation") an approach introduced by Warfield [2004] in which a probabilistic estimate of the true segmentation is computed. Here, the two approaches produce nearly identical

TABLE II. Average confusion matrices, BrainWeb variable quality data

		gm	wm	om
SPM5	gm	0.891 ± 0.052	0.068 ± 0.047	0.041 ± 0.017
FreeSurfer		0.803 ± 0.011	0.129 ± 0.012	0.068 ± 0.002
FSL		0.828 ± 0.019	0.069 ± 0.028	0.104 ± 0.013
SPM5	wm	0.091 ± 0.028	0.908 ± 0.028	0.005 ± 0.002
FreeSurfer		0.092 ± 0.007	0.906 ± 0.007	0.002 ± 0.000
FSL		0.130 ± 0.058	0.868 ± 0.058	0.002 ± 0.000
SPM5	om	0.008 ± 0.003	0.000 ± 0.000	0.991 ± 0.003
FreeSurfer		0.005 ± 0.000	0.000 ± 0.000	0.995 ± 0.000
FSL		0.009 ± 0.002	0.000 ± 0.000	0.991 ± 0.002

Values are the arithmetic means of the confusion matrices of all data sets ± standard deviation, gm: gray; wm: white; and om: other (all non-gm and non-wm) matter.

results. Except for the fact that SPM5 and FSL alone produce slightly less false positive voxels (SPM5-WM and FSL-GM/WM), the results obtained with the two fusing methods overall are as good as or even slightly better than the best single-segmenter segmentation in comparison with the reference values. These results derived from a single-slice example are supported by the analysis of a larger dataset that confirms that STAPLE and VOTING produce the same results in nearly all cases and that they perform equally well as the best single segmenter (SPM for gm, SPM and FSL for wm), but do not significantly improve the segmentation accuracy (see supplementary Table I). The results also show that FreeSurfer is comparatively bad at correctly classifying gray and white matter for all evaluated datasets (see also "confusion matrices").

The reason for the quasi-identity of the two methods might be the high similarity of the SPM5 and FSL segmentation results as compared with FreeSurfer. The advantages of the STAPLE approach as described by Warfield et al. [2004] might be more obvious in situations in which more segmenters are used and the results are more incongruent.

Variable Quality BrainWeb Data

SPM5

The deviations of the gray matter classification results from the reference value ranged between −3.9 and +1.9%, μ[3] = −1.6 ± 1.8% (±standard deviation) with coregister preprocessing[4] (without: −5.2 and +5.0%, μ = 1.0 ± 3.8%) for different image qualities, white matter results ranged between −9.6 and +8.3%, μ = −0.01 ± 6.1%. Since these deviations partially counterbalance each other, total brain matter volume showed smaller variations between −3.1 and +1.3%, μ = −1.2 ± 1.9% (Figs. 4 and 5).

Since registration-transformed data sets lead to more accurate results than un-preprocessed data sets, we preprocessed all other data sets with the coregister function in this study.

Correlating the deviations with the data quality, a tendency can be seen that gray matter volumes are overestimated for good quality data and underestimated for data of relatively bad quality and vice versa for white matter. Quantification of this observation led to the Spearman rank correlation coefficient $r_s = 0.94$ for white and $r_s = -0.62$ for gray matter (significance levels $P < 0.002$).

FreeSurfer

FreeSurfer underestimates gray matter volumes for all image qualities and the deviations ranged between −5.5 and −8.0% (μ = −6.5 ± 0.8%). In contrast to gray matter, white matter volume results are too high in FreeSurfer (μ = 10.0 ± 1.5%), so that the GM and WM sums yielded values relatively close to the reference values (0.0 to 0.8%, μ = 0.6 ± 0.3%), only slightly overestimating total volumes. In FreeSurfer all deviations have the same sign within each group.

No correlation between image quality and segmentation performance could be observed (Spearman rank correlation $|r_s| < 0.15$ for white and gray matter), and even though FreeSurfer shows relatively high deviations of single white or gray matter from the reference values it can be considered the most consistent method in terms of dependency of deviation on image quality. This is because the deviations of white and gray matter from the reference values have standard deviations of 1.5 and 0.8%, much less than for SPM5 (WM: 6.1 and GM: 3.8%) and FSL (WM: 9.4/6.1 and GM: 5.8/3.8%, with/without data set 2).

FSL

FSL (PBMAP) underestimates all total brain volumes, 8 out of 10 gray matter volumes and 6 out of 10 white matter volumes. When leaving out the extreme values for data set n1rf40[5], gray matter volumes vary between data sets of different quality from −6.6 to +0.1% (μ = −4.6 ± 2.1%) and white matter volumes from −10.6 to +5.4% (μ = −1.2 ± 5.3%). Total volumes show values between −5.1 and −1.2% (μ = −3.3 ± 1.4%). Similar to SPM5, a correlation between ranked data quality and deviation can be detected for white (Spearman $r_s = 0.88$, $P < 0.05$), but not for gray matter, and the deviations do not show a stable pattern similar to that present in the FreeSurfer data.

For APRIORI we find the following results: in contrast to PBMAP, APRIORI consistently overestimates gray μ =

[3] μ is an estimate of the arithmetic mean. It has to be noted that the arithmetic mean of the deviations of the computed from the reference volumes may in certain cases be misleading if negative and positive deviations counterbalance each other and result in small average deviations suggesting a high accuracy. Thus it should always be used in connection with the standard deviation as in the above statistics, which gives a good estimate of the variance of the deviations.

[4] We present two variations of SPM5 results: with and without (shown in brackets) rigid body transformation using the SPM5 function coregister, for details see Materials and Methods section.

[5] We excluded this extreme outlier from the analysis to ensure nonexaggerated statistical results.

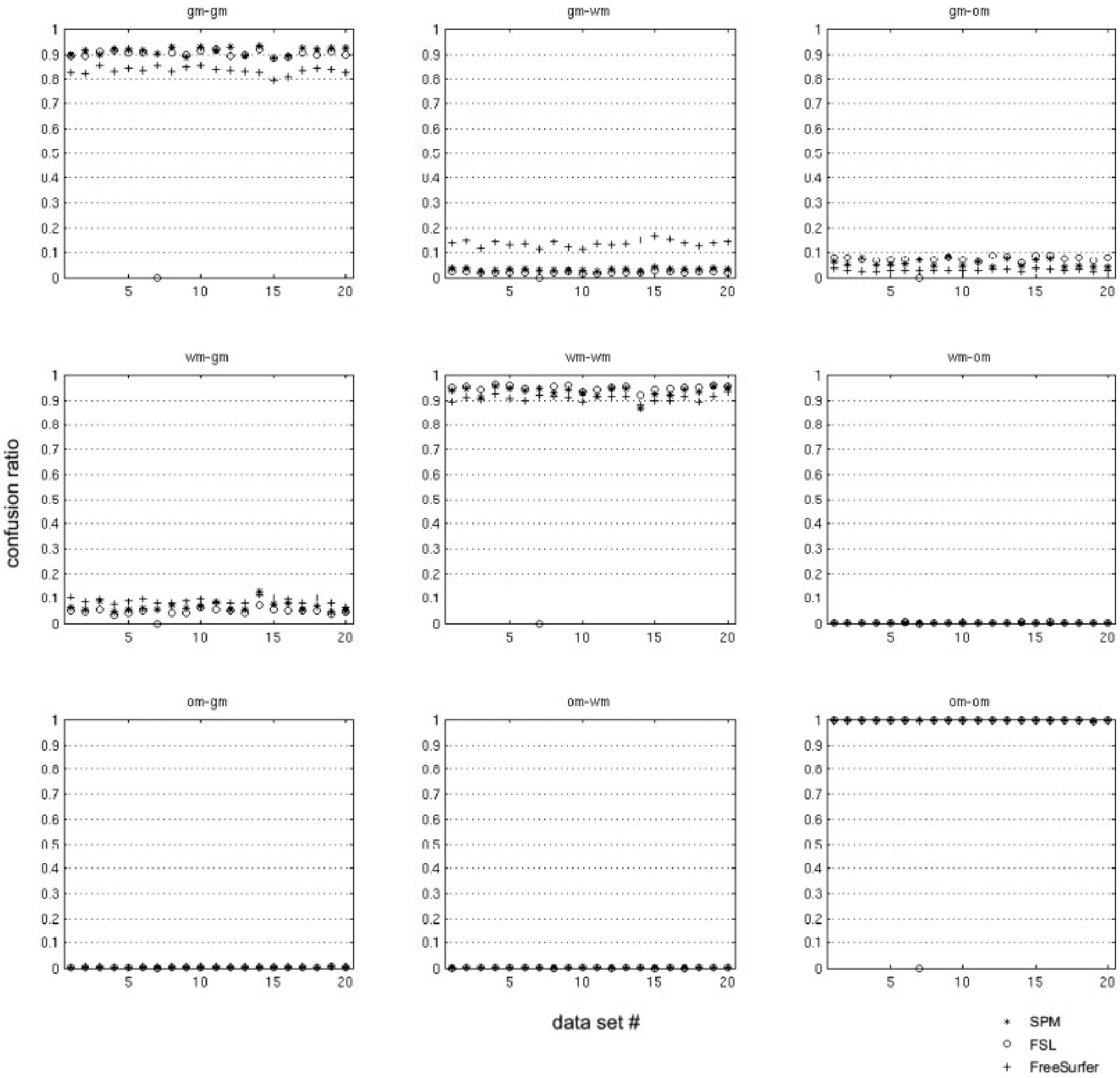

Figure 2.
Multi-confusion-matrix (multiple-anatomical model BrainWeb data): Each element of the confusion matrix contains the results for the three segmenters for all 20 data sets (compare Fig. 1).

6.0 ± 5.4% (range 2.4 to 16%) and underestimates white matter −7.9% ± 8.7% (range −23 to 0.8%). PVE has average deviations of −1.2% ± 4.9% (range −4.5 to 12%) for gray and 0.7% ± 7.5% (range −18.3 to +7.5%).

Optimization of volume counts by data fusion (VOTING)

We also investigated if data fusion by VOTING improves accuracy and stability of volume counts (we only perform VOTING because of the quasi-identity of the STAPLE/VOTING results presented in section "Optimization by Data Fusion"). We tested different VOTING scenarios (two or three out of four SPM, FreeSurfer, FSL pbmap, FSL a priori) and also found no significant improvement as compared with single methods. Neither the accuracy nor the stability increased in VOTING results (average of scenario deviation means: μ = 5.3% ± 1.8% (stability: range = 7.4%) for gray and μ = 2.3% ± 1.5% (range = 7%) for white matter).

TABLE III. Average confusion matrices, multiple-anatomical-model BrainWeb data

		gm	wm	om
SPM5	gm	0.913 ± 0.015	0.031 ± 0.005	0.056 ± 0.015
FreeSurfer		0.832 ± 0.015	0.138 ± 0.014	0.029 ± 0.004
FSL		0.904 ± 0.010	0.020 ± 0.003	0.076 ± 0.009
SPM5	wm	0.065 ± 0.019	0.933 ± 0.020	0.014 ± 0.011
FreeSurfer		0.089 ± 0.013	0.908 ± 0.013	0.029 ± 0.001
FSL		0.048 ± 0.009	0.949 ± 0.010	0.003 ± 0.002
SPM5	om	0.001 ± 0.001	0.000 ± 0.000	0.999 ± 0.001
FreeSurfer		0.005 ± 0.001	0.000 ± 0.000	0.995 ± 0.001
FSL		0.000 ± 0.001	0.000 ± 0.000	1.000 ± 0.001

Values are the arithmetic means of the confusion matrices of all data sets ± standard deviation, gm: gray, wm: white and om: other (all non-gm and non-wm) matter.

Multiple Anatomical Model BrainWeb Data

SPM5

For the multiple-anatomical-model BrainWeb data (see Fig. 6), gray matter results were underestimated in 18 of 20 cases with the deviations ranging from −7 to +3%, μ = −3.5% ± 2.5%. The situation for white is not as clear as for gray matter, but WM volumes are underestimated by SPM5 on average as well ranging from −10 to +3%, μ = −1.6% ± 2.9%.

Total brain parenchyma volume deviations between −5.0 and 1.7% result in an on average underestimation of gray plus white matter of μ = −2.8% ± −2.8%.

FreeSurfer

For the multiple-anatomical-model BrainWeb data sets, FreeSurfer always underestimated gray (μ = −5.0% ± 1.9%) and always overestimated white matter results (μ = 12.8% ± 3.4%). Deviations from the reference value for white matter ranged from 6.7 to 19.6% and from −1.8 to 8.0% for gray matter. These deviations partially counterbalance each other so that total brain parenchyma volumes are more accurate (μ = 1.8% ± 1.1%) ranging from 0.01 to 5.6%.

FSL

The comparison of the FSL segmentation results obtained with the probability map (PBMAP), partial volume estimation (PVE), and a priori information (APRIORI) methods shows the following differences. APRIORI segmentation on average underestimates gray matter by μ = −1.5% ± 1.7% (range: −4.1 to +3.3%), which is slightly better than PVE (μ = −1.8% ± 2.0%) (range: −4.2 to +3.9%) and less underestimation than PBMAP (−6.4% ± 1.7%) (range: −8.2 to −0.9%).

For white matter the situation is nearly opposite, APRIORI show stronger underestimation (μ = −5.3% ± 1.3%) (range −8.1 to −2.9%) than PBMAP (μ = −1.7% ± 1.1%) (range −4.8 to +0.6%), whereas PVE overestimates white matter (μ = 2.0% ± 1.1%). See also Table V.

Evaluation of Segmenter-Specific Aspects

Coregistration improves results in SPM5

Segmentation of the variable-quality BrainWeb data showed that the rigid registration "coregister" function of SPM5 improved the segmentation accuracy (see Figs. 1, 4, and 5). We therefore used coregistration for all data sets by default unless otherwise noted.

Cortical gray matter: Comparison of surface- and volume-based results in FreeSurfer

FreeSurfer offers two different approaches to calculate cortical gray matter (CGM) volumes: One is based on the surface triangulation and cortical thickness computed for each vertex (surface-based) and the other computes the volumes by summing up all cortical gray matter voxels (volume-based).

We compared the results of the two methods for the nine BrainWeb data sets of variable image quality. The surface-based results consistently show smaller CGM volumes than the cortical volume-based results ranging from 9.5 to 14.4% (μ = 13.1% ± 1.6%). The BrainWeb reference model only provides a single class label for gray matter. It is thus not possible to exactly determine the accuracy of the CGM results. In case of total gray matter, FreeSurfer consistently underestimates volumes for these BrainWeb data sets. If the assumption is made that this underestimation is equally distributed percentagewise between cortical and subcortical gray matter (which seems plausible, but cannot be proven here), this would imply that the volume-based is more accurate than the surface-based method.

Impact of different brain extraction methods

In FreeSurfer and SPM5, the brain extraction/masking process is integrated into the brain tissue segmentation procedure, whereas in FSL the segmentation procedure is split into segmentation of brain/non-brain-tissue (brain extraction/masking) and segmentation of brain tissue into white and gray matter and CSF. In FSL, if a generously sized mask is used or the brain extraction preprocessing is omitted completely, the segmentation is highly erroneous because FAST cannot distinguish between brain and non-brain tissues leading to huge errors of more than 100%

TABLE IV. Gray and white matter sensitivities and specificities based on BrainWeb data

	SPM5	FreeSurfer	FSL
gm-sensitivity	0.90	0.82	0.87
wm-sensitivity	0.91	0.90	0.91
gm-specificity	0.96	0.95	0.95
wm-specificity	0.97	0.93	0.98

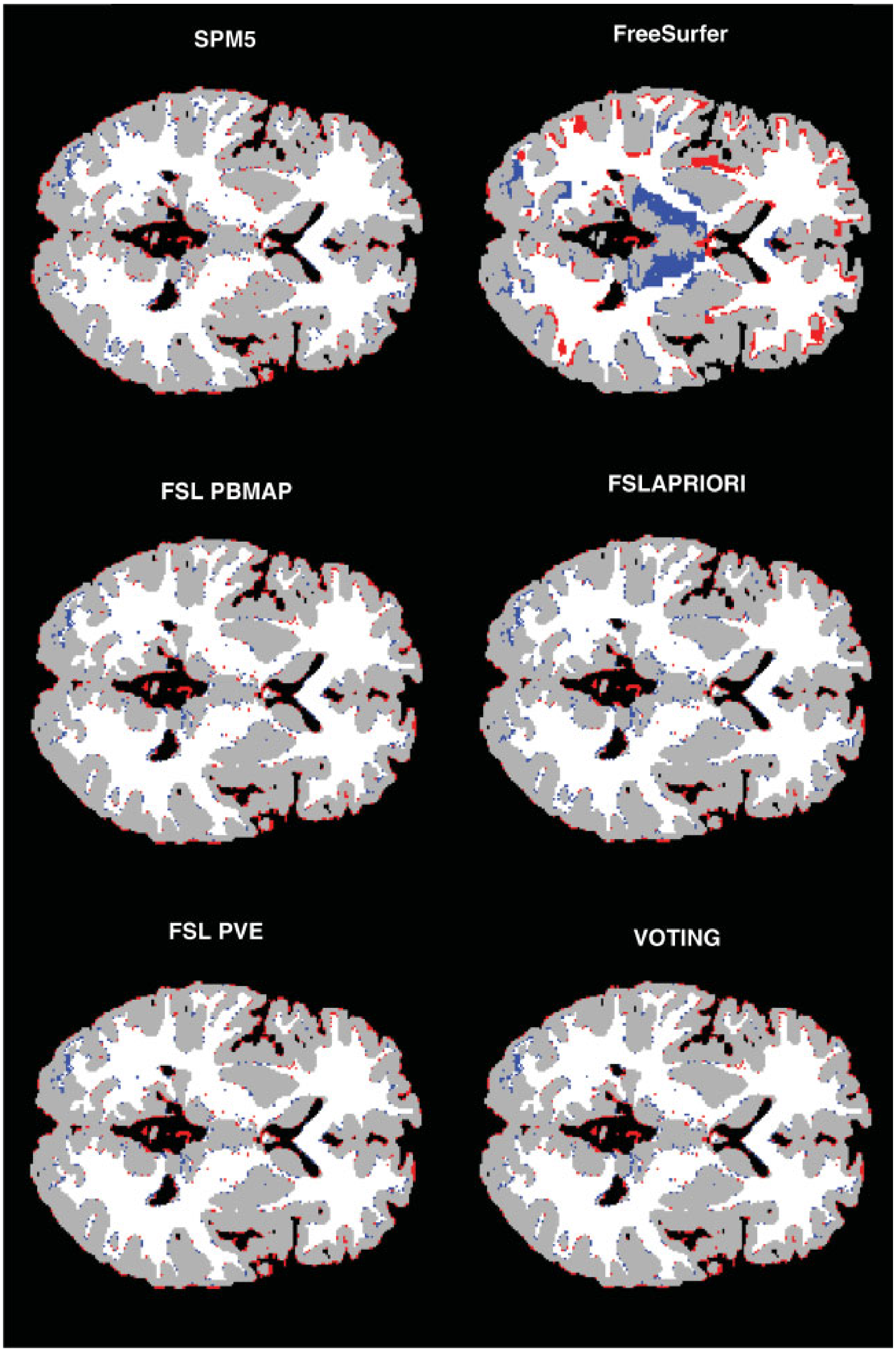

Figure 3.

Multiple-anatomical-model BrainWeb data set 06 slice 73. Results for different methods and VOTING overlaid with the template-reference data. Red/blue voxels: segmentation does not agree with reference data (red: gray matter; blue: white matter).

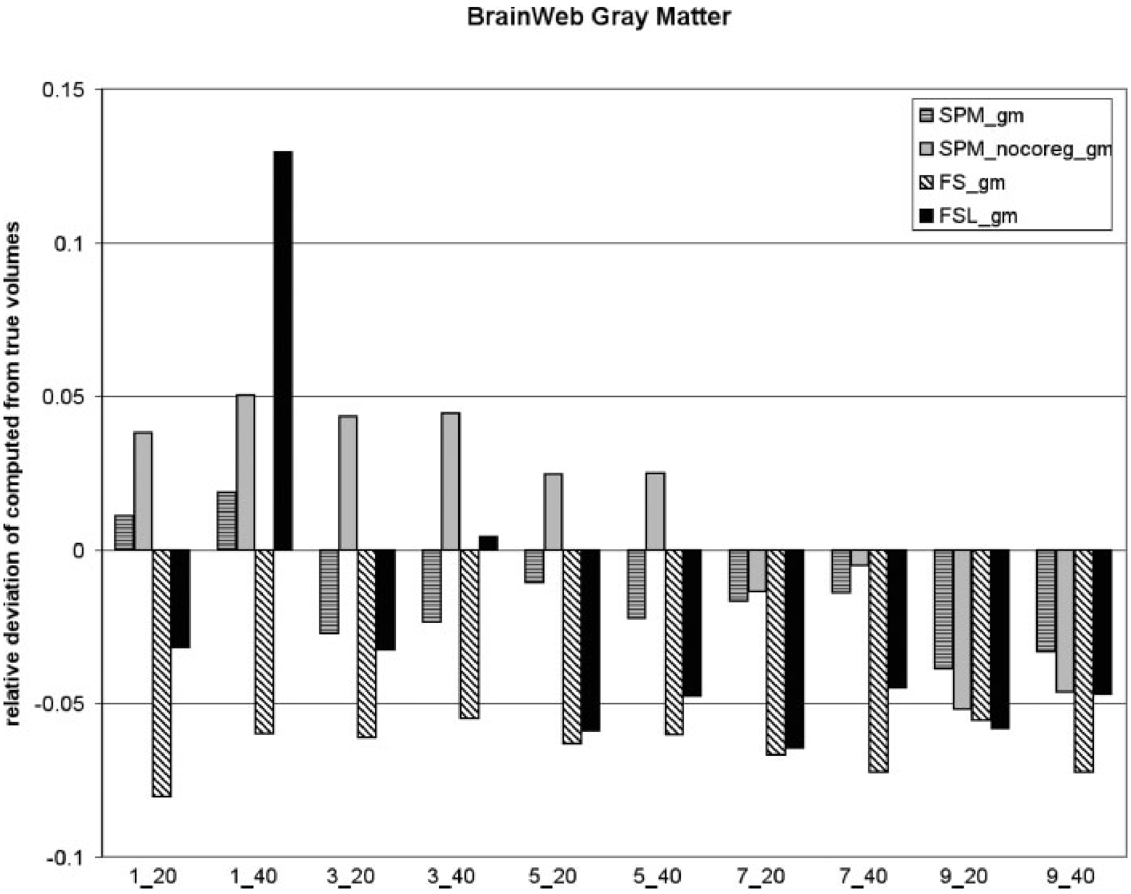

Figure 4.

Comparison of gray matter results for varying image qualities based on same template brain. The two different values given for SPM5 (no suffix/suffix_nocoreg) denote the different results obtained with/without coregistration option.

overestimation as compared with the segmentation results after brain extraction (results not shown).

To evaluate the impact of the brain masking/extraction procedure on the segmentation results, we processed the data sets with FSL using BET and BSE [Fennema-Notestine et al., 2006]. We then compared the segmentation results obtained after BET or BSE preprocessing. We visually inspected the processed images that showed deviations between BET/BSE-preprocessed segmentation results of more than 5% to check whether the differences were caused by an incorrect brain extraction process. For BrainWeb, data sets 4 and 41 showed large parts of the calvarium that were not deleted. Excluding these two, the data sets show discrepancies between BET- and BSE-based segmentations (using the same segmentation parameters) of mean = 0.2% ± 0.3% (median = 0.1%) for white and mean = 0.5% ± 1% (median = 0.3%) for gray matter, which is nearly one order of magnitude smaller than the discrepancies between segmentation results due to different segmentation methods (average of BrainWeb deviations: mean = 3.1% ± 3.0% for gray and 1.9% ± 3.3% for white matter results of segmentations using probability maps, partial volume estimation, and a priori information in FAST/FSL). We also ran two separate analyses of the OASIS data sets with BSE and BET and the results support the earlier findings for BrainWeb. From these results we can conclude that first, discrepant segmentation results are caused mainly by differences in the brain-tissue (gray/white matter) segmentation capabilities of the particular methods and second, that the different masking methods BSE and BET show a negligible effect on the segmentation result provided that the brain extraction process did not fail/did not show any strong, visually prominent errors.

Comparison of Real Data Results

The analysis of the real data sets shows discrepancies between different segmenters of 13.2% ± 10.2% on average (maximum 36.6%) for white and μ = 13.9% ± 7.6% (up to 41.6%) for gray matter based on the ratio of the minimal

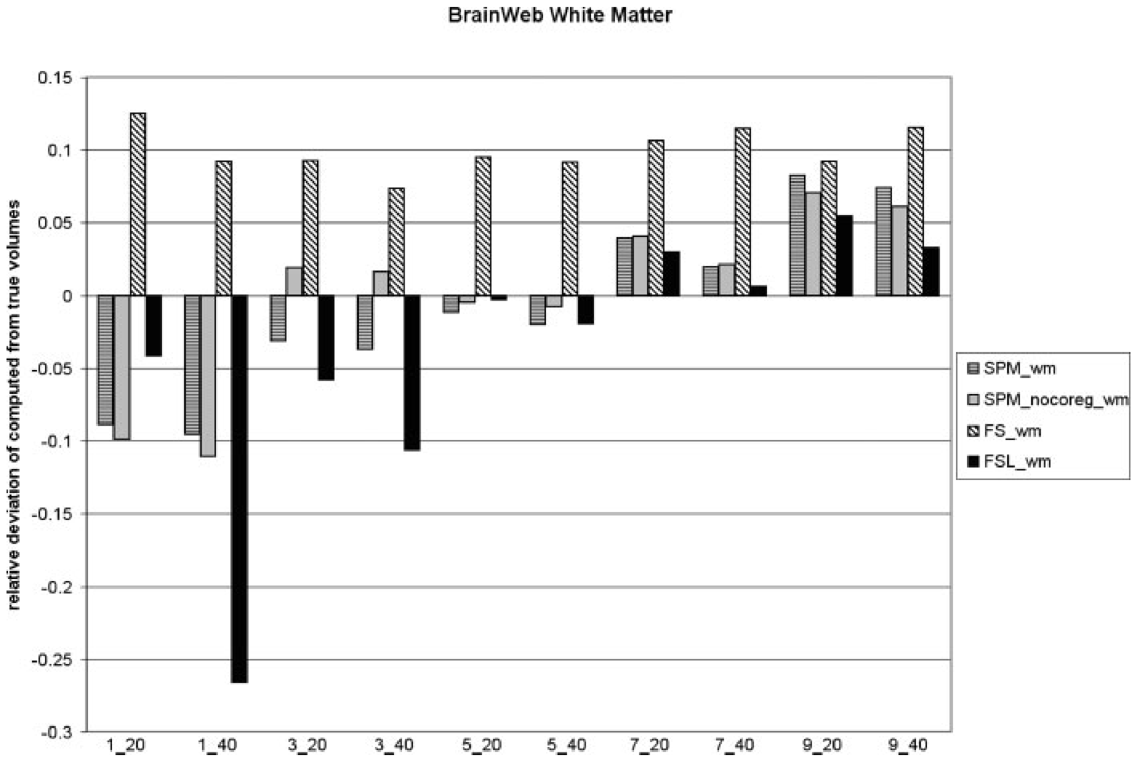

Figure 5.

Comparison of white matter results for varying image qualities based on same template brain. The two different values given for SPM5 (no suffix/suffix_nocoreg) denote the different results obtained with/without coregistration option.

and maximal volume for each data set. Initially, following the recommendation for the default procedure with SPM5, we did not use the coregister function (see Materials and Methods section) to preprocess the data sets, which resulted in gray matter volumes about 50% higher in FSL or FreeSurfer than in SPM5. Preprocessing with the coregister function significantly decreased these differences (SPM5 gray matter volumes increased by 34% on average for the data sets 501, 556, 558). We thus decided to always coregister-preprocess the data when using SPM5 and not to include the unpreprocessed results in the global statistical analysis.

For white matter, FreeSurfer on average computed the highest (~8% higher than SPM5) and SPM5 the lowest volumes (~3% lower than FSL). FSL on average takes a middle position. Conversely, FreeSurfer computes the lowest gray matter volumes (~9% lower than FSL). However, for gray matter, FSL-derived volumes are highest on average, but only about 2% deviant from SPM. Although for GM results no significant differences between the MP-RAGE and other data sets can be found, the white matter MP-RAGE results show higher average maximum discrepancies ($\mu = 30.7\% \pm 8.1\%$) than the rest of the data ($\mu = 9.4\% \pm 5.3\%$). These results could hint at a possible influence of the acquisition technique on segmentation results in certain cases.

Supplementary Figure 3 shows a visualization example of subject 560 slice 62 similar to that presented for the BrainWeb data. However, since no reference data set is available in this case, segmentation results were compared pairwise (FreeSurfer-SPM5, FreeSurfer-FSL, and SPM5-FSL). The overlaid representations of FreeSurfer-SPM5 and FreeSurfer-FSL show a similar pattern: FreeSurfer consistently underestimates gray matter and overestimates white matter volumes when compared with both SPM5 and FSL segmentations, thus calculating higher white and lower gray matter volumes than both SPM5 and FSL. SPM5 and FSL segmentations show a relatively higher consensus especially for gray matter (ratio of number of identical voxels to number of total voxels: 0.67, for details see supplementary Table III).

These findings show some analogy to the results for the BrainWeb data, where SPM5 and FSL segmentations were also similar in comparison with FreeSurfer.

Intra-individual inter-timepoint segmentation differences

A highly relevant question is whether MR scans of the same individual at different time points using the same scanner, protocols, and software can lead to differences in

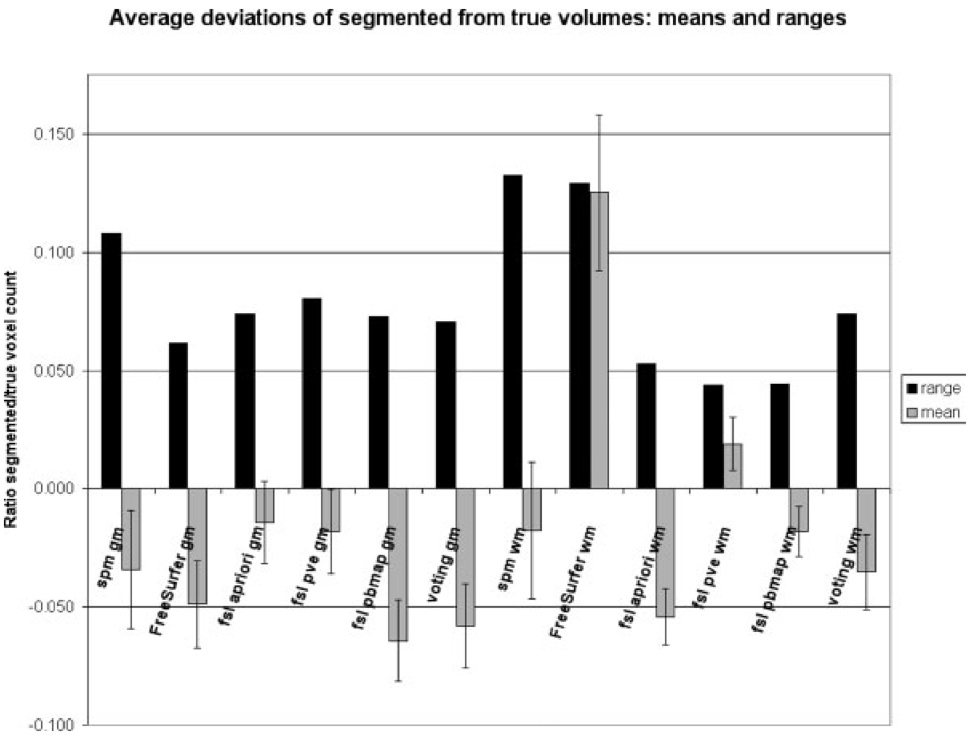

Figure 6.
Means (with standard deviations) and ranges of deviations of segmented from true volumes for BrainWeb data (multiple-anatomical-model), gray (gm), and white (wm) matter. Values are averaged over all data sets for each method. Ranges are differences between largest and smallest (highest positive and highest negative) deviations.

the segmentation results. We analyzed 48 pairs of datasets provided by the OASIS database that were acquired from individuals at two time points not more than 90 days apart. Since ageing-related changes occur on longer timescales and only subjects without pathological changes are included in this group, comparing the segmentation results may help to investigate to what extent volumetry differences can be caused by effects [noise, movement artifacts, (slightly) different position in scanner, temperature, . . .] not related to an actual change of the imaged brain tissue. These data can be considered as the "real" counterpart to the simulated BrainWeb data sets used in this study that provide a means to assess brain volumetry robustness of anatomically identical brains with varying image quality.

Our analysis shows that differences between two scans do exist. Although FreeSurfer on average computes volumes that differ ~1.5% for the same subject for both white and gray matter and FSL has a similar behavior for gray matter, the white matter inter-scan deviations of FSL have a mean of over 4%. SPM5 computes more reliable results deviating only around 0.7% for white and gray matter. However, these numbers present estimates of the mean difference and even for SPM differences of over 3% occur while the other segmenters show maximum discrepancies of 4.4% (FreeSurfer gm) and 5.2% (FSL gm) and 10.3% (FSL wm).

In patients for whom MR scans over time are used to evaluate the progression of disease or efficacy of a treatment, it has to be taken into careful consideration that the effects already seen in the BrainWeb data and confirmed here using real image datasets potentially over- or understate pathological changes of brain volume. In this context, our results suggest that SPM5 might be more reliable than the other segmenters.

Impact of ageing on segmentation results

Closely related to what has been described in the previous section, we used OASIS brain data from different age

TABLE V. Multiple-anatomical model BrainWeb deviation of segmentation results from references values

μ ± SD (%)	Gray matter	White matter
SPM5	−3.5 ± 2.5	−1.6 ± 2.9
FreeSurfer	−5.0 ± 1.9	+12.8 ± 3.4
FSL PBMAP	−6.4 ± 1.7	−1.7 ± 1.1
FSL PVE	+1.8 ± 2.0	+2.0 ± 1.1
FSL APRIORI	−1.5 ± 1.7	−5.3 ± 1.3

groups to investigate whether relative discrepancies of the volumetry results between segmenters show a dependency on age. We analyzed the segmentations results for a group of young (age <35 yrs, 13 individuals) and old (>69 yrs, 16 individuals). The results show that FreeSurfer underestimates gray matter when compared with both SPM and FSL more in the young than in the old group (gm counts ratio FS/SPM: μ(young) = $-11.0\% \pm 3.7\%$, μ(old) = $0.3\% \pm 9.2\%$, t-test—significance of difference $P = 0.0003$), FS/FSL: μ(young) = $-14.0\% \pm 6.7\%$, μ(old) = $-6.4\% \pm 7.4\%$, t-test—significance of difference $P = 0.007$). The comparison of the FSL/SPM ratios did not show statistically significant differences. Given the relatively small sample size, drawing conclusions from this finding has to be done carefully, especially because our observation is counter-intuitive because of a (normal) loss of gray matter tissue with age. The effect might relate to age dependent signal intensities of tissue or the effect of partial voluming on the different segmenters as with age there might be slightly more CSF in contact with slightly less parenchyma. Another issue is that the age effect could be related to differences in the a priori model and templates. Further investigation of this issue is important especially with regard to ageing and life span studies in which brain volumetry/morphometry are linked to cognition and genotype.

Error Analysis—Estimation of Required Number of Subjects

To what extent do the segmenter-dependent effects described in the previous sections influence the planning and analysis of clinical studies? Providing concrete and general advice is difficult because it would depend on several quantitative unknown factors, such as the size of the suspected difference and the quality of the data in general and systematic quality issues with regard to disease specific effects (e.g., motion artifacts in patients with dementia). As we have shown, the segmentations results depend on the data quality (noise, intensity-inhomogeneities). However, we could show that (except for some aspects of SPM5 results) no obvious correlation between data quality and segmentation results can be derived. This means that even if quality parameters could be easily and robustly measured or estimated from imaging data, deriving a corrective factor would remain difficult if not impossible. To address this issue and provide some guidance to the readers, we performed Monte-Carlo simulations based on the multiple-anatomical-model BrainWeb data, by creating a two data sets: (1) the original BrainWeb data (numbers increased to 100, 120, and 150 by randomly generating gray matter counts in the range given by the minimal/maximal values in the BrainWeb data) and (2) a group that was created by multiplying each of the datasets in the first group with a gray matter reducing "disease-factor" (randomly generated up to 3% negative deviation) and a "segmenter-factor" (randomly generated up to 5% deviation). This would model the situation in a study in which a group of patients is examined at two time points, e.g., to assess disease progression). We then performed Student's t-tests to assess the impact of the segmenter-factor on the significance level. We ran 100 Monte-Carlo simulations each for $n = 100$, 120, and 150 subjects and obtained the following mean t-values (disease-factor + segmenter-factor/disease-factor alone): $n = 100$: ($t = 1.7/t = 2.1$); $n = 120$: ($t = 2.0/t = 2.3$); $n = 150$: ($t = 2.3/t = 2.6$).

These simulation results show that the segmenter-factor reduces the t-values (and thus the significance level) by ~0.3. When a 98% confidence of significance is desired, $n = 150$ subjects are necessary to obtain the required t-value when the simulation includes the segmenter-factor, i.e., the situation we observe in the results presented in this study. In our simulations, the same significance level could be obtained with 120 subjects only if the segmentation process would be perfectly accurate.

This example illustrates how clinical investigators could estimate the minimal number of subjects needed if an assumption about the expected volume differences can be made. In the Monte-Carlo simulation presented here, segmenter-dependent effects require an increase of the number of subjects by ~25%.

In contrast to the segmenter-dependent effects on group comparisons that can be handled as shown here, our analysis suggests that detecting intra-individual time-course differences in the order of one-digit percentages (i.e., patients scanned at different time points to evaluate disease progression) does not produce meaningful results.

DISCUSSION

Our analysis of real and simulated data sets uncovered pronounced within- and between-method variation in segmentation results for whole brain and total gray and white matter volume. In SPM5, the results for the variable-quality BrainWeb data suggest the following correlation between volumetric accuracy and image quality: white/gray matter volumes tend to be under-/overestimated for good quality data and over-/underestimated for data of relatively bad quality, with the smallest deviations for mediocre data. A similar observation can be made for white matter FSL results. Moreover, the confusion matrix analysis of the variable quality BrainWeb data reveals a positive correlation between SPM5 gray matter detection sensitivity and increasing data quality. FreeSurfer consistently under-(over-)estimated gray (white) matter volume for all BrainWeb data sets. On the other hand, it was relatively accurate for brain parenchyma (i.e. GM + WM) volumetry and FreeSurfer was more robust than FSL and SPM5 to changes in image quality and our data suggest that it is more suitable for GM + WM volume measurements than the other segmenters when image quality is indeterminable or varies within a study. The confusion matrix analysis suggests that mainly SPM5 and also FSL are more suitable than FreeSurfer for tasks where high gray matter sensitivity is required, i.e., high confidence that voxels belong to gray matter are classified as gray matter (see Fig. 3). In this case, the best performance is achieved when using SPM5 with high-quality data (i.e., noise level up to ~3%). These results presented in the "gm"-row of the variable-quality confusion matrix (see Fig. 1) may also hint at a higher contribution of noise in comparison with intensity-inhomogeneity to the image-quality-related instability of SPM5: pairs of data sets with the same noise level and different intensity-inhomogeneities (20 or 40%) show practically the same confusion matrix value, while differences are obvious when increasing noise and keeping intensity-inhomogeneity levels constant, a phenomenon also visible in the SPM5 volume counts (see Figs. 4 and 5). This is particularly interesting with regard to 3 Tesla MR imaging in which intensity-inhomogeneities are more pronounced than at lower field strengths (here we deal with 1.5T data). Future studies e. g. comparing individuals scanned both at 1.5 and 3T or simulated 3T BrainWeb data sets based on the same atlas as the present 1.5T data will be necessary to investigate these issues.

Altogether it can be said for the variable-quality BrainWeb data that with deviations from the reference values of the volumetry results of more than 10% for SPM5, FreeSurfer and FSL even using data of relatively good quality, the accuracy of the methods must be considered as relatively low. Thus, comparing volumetry results obtained with different methods is not advisable. The comparison, for example, of results for white matter calculated by SPM5 with those produced by FreeSurfer, can yield discrepancies of 20%. The analysis of the multiple-anatomical model BrainWeb data shows similar results. Both arithmetic mean and standard-deviation/range of the deviations of the segmented from reference values (measures to estimate accuracy and stability of the methods) show percentage values in the order of single digits up to low two digits. On average, gray and white matter are underestimated with all methods. The only strong exception is FreeSurfer that shows an overestimation of over 10% on average. Stability is best with FSL PBMAP/PVE for white (deviations <5%) and FreeSurfer is the best method to stably compute gray matter volumes (range = 6.2%) followed by FSL PBMAP and APRIORI (range 7.3 and 7.4%). SPM gray and white and FreeSurfer white matter volume counts even show ranges of over 10%.

These conclusions drawn from the analysis of simulated data are supported by the findings that the real data results show even higher variations between segmentations. These strong variations also make clear how difficult it is to confidently estimate the actual brain volume from the segmented results. Even though for the 57 real and 20 multiple-anatomical-model BrainWeb data sets tested here, the differences/deviations show some qualitative similarities, it would be difficult to deduce a segmenter-specific quantitative factor to "normalize" the data to make inter-segmenter comparison possible because of relatively high variations of the discrepancies.

In addition to that and even more importantly, the low stability of the methods shows that brain volumes determined with the same method should be considered carefully when compared, because relatively small differences in image quality can have an impact on the deviations of the calculated volumes from the "true" volumes that do not follow any regular pattern. This finding is particularly significant for the results for white matter using SPM5, where the volumes of two different image qualities differ over 17%, or for the values for the gray matter volumes (discrepancy over 10%), but is present nearly throughout the volumetry results. These results obtained from the variable-quality BrainWeb data are supported by our findings that segmentations of (real) data from healthy volunteers (OASIS data base) scanned twice within weeks show similar discrepancies between time points even when using the same method and constant image acquisition conditions.

In this context, the study by Chard et al. [2002] about SPM99 volumetry reproducibility over time shows the remarkable result that coefficients of variation ranged from 1.2 to 0.5% for individuals undergoing subsequent MR scans 197 days apart. According to our results such relatively small variations between segmentations of subsequent scans suggest that the data analyzed by Chard et al. [2002] must fulfill higher quality standards than the data we analyzed.

These within-segmenter comparisons have been performed using simulated images, because this allows to control and change specific quality parameters for the same original data set. In this study, the common quality parameters "noise" and "intensity-non-uniformity" were varied over typically occurring ranges [Collins et al., 1998; Kwan et al., 1999] for the variable-quality BrainWeb data, whereas the multiple-anatomical-model data has constant image quality/acquisition parameters and varies the anatomical features.

In real MRI data, more parameters than those simulated here do vary, i.e., the simulation does not include parameters like motion artifacts or scanner-specific technical characteristics, which can have a high impact on image quality. This "limitation" of the simulated data, however, does not invalidate the results and conclusions presented, but rather suggests that the effects described here are even more pronounced in real data sets, an assumption that is supported by the much larger between-segmenter differences for the tested real than for simulated data.

The relevance of these results is based on the impact they might have on clinical applications such as detecting relatively small but diagnostically relevant changes in patients over time and studies using these methods for brain volumetry to investigate connections between diseases and volumetric changes [e.g., Job et al. 2002], because differences in image quality linked to the discussed effects occur in serial MR-examinations even if the same scanner and protocol are used (see our analysis of OASIS data at two different time points). In studies using data acquired with different MR-scanners or different protocols, variations of image quality are likely even more frequent.

Since normal brain volume shows a significant variation anyway (in the multiple-anatomical-model BrainWeb data the biggest gray matter volume is 24% larger than the smallest (30% for white matter) and the expected volume differences due to pathological conditions are usually rather subtle, the instability of the segmentation methods has a significant influence on statistical analysis of group comparisons. Our Monte-Carlo simulations show an intuitive method to estimate the necessary increase in the number of subjects in a study due to the segmenter-instability. The longitudinal study model we test shows that 150 instead of 120 patients are needed to detect a gray matter volume decrease of up to 3% between two time points. Comparing groups of patients is likely to even make the increase higher.

In this study, we used standard procedures with default parameters as suggested in the documentations of the software packages, because we wanted to assess the performance of the segmenters according to the way they are used by most (applied) researchers. Exceptions are the use of the coregistration function in SPM5, which probably enhances the use of a priori information through an improved alignment of the image data with the template brain and the additional use of partial volume estimation and a priori information in FSL. Especially the FSL analysis shows how difficult it is to give a general recommendation about specific segmentation methods. Given the relatively low stability (ranges of over 10% for the multiple-anatomical-model data and up to over 20% for the variable-quality BrainWeb data) of the segmentations the mean values estimating the accuracy should not be overinterpreted to give definite advice which parameter settings to prefer even if the differences between some mean deviation values would suggest that. An example is that the multiple-anatomical-model data suggests that APRIORI performs better for gray matter segmentation, whereas PBMAP is more accurate for white matter. However, when comparing the results for the variable quality BrainWeb data APRIORI even shows a slightly lower gray matter accuracy than PBMAP.

The positive impact of the usage of the SPM5-coregister function on segmentation results, the tests of fusing segmentation results performed here and the differences between PBMAP, PVE and APRIORI segmentation in FSL or between the volume- and surface-based approach in FreeSurfer exemplify how future studies about systematic parameter variation and combination of different methods could help optimize segmenter performance. Unfortunately, the optimization approaches STAPLE/VOTING we tested do not seem to present a straight-forward solution of the shortcomings of the methods that we uncovered in this study. The general conclusions of our study remain valid both for single and fusion-methods.

ACKNOWLEDGMENTS

A part of the real data sets were acquired at the Haraldsplass Deaconess University Hospital, Bergen, Norway, as part of a larger study of aging supported by the Research Council of Norway and Kavli Center for Dementia Research.

REFERENCES

Amato U, Larobina M, Antoniadis A, Alfano B (2003): Segmentation of magnetic resonance brain images through discriminant analysis. J Neurosci Meth 131:65–74.

Ashburner J, Friston KJ (2005): Unified segmentation. NeuroImage. 26:839–851.

Aubert-Broche B, Collins DL, Evans AC (2006): A new improved version of the realistic digital brain phantom. Neuroimage 32:138–145.

Aubert-Broche B, Griffin M, Pike GB, Evans AC, Collins DL (2006): 20 new digital brain phantoms for creation of validation image data bases. IEEE TMI 25:1410–1416.

Barnes J, Whitwell JL, Frost C, Josephs KA, Rossor M, Fox NC (2006): Measurements of the amygdala and hippocampus in pathologically confirmed Alzheimer disease and frontotemporal lobar degeneration. Arch Neurol 63:1434–1439.

Barra V, Boire JY (2001): Automatic segmentation of subcortical brain structures in MR images using information fusion. IEEE Trans Med Imaging 20:549–558.

Bezdek JC, Hall LO, Clarke LP (1993): Review of MR image segmentation techniques using pattern recognition. Med Phys 20:1033–1048.

Byrum CE, MacFall JR, Charles HC, Chitilla VR, Boyko OB, Upchurch L, Smith JS, Rajagopalan P, Passe R, Kim D, Xanthakos S, Ranga K, Krishnan R (1996): Accuracy and reproducibility of brain and tissue volumes using a magnetic resonance segmentation method. Psychiatry Res 67:215–234.

Cates JE, Lefohn AE, Whitaker RT (2004): GIST: An interactive, GPU-based level set segmentation tool for 3D medical images. Med Image Anal 8:217–231.

Ciumas C, Savic I (2006): Structural changes in patients with primary generalized tonic and clonic seizures. Neurology 67:683–686.

Chard DT, Parker GJ, Griffin CM, Thompson AJ, Miller DH (2002): The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. J Magn Reson Imaging 15:259–267.

Clark KA, Woods RP, Rottenberg DA, Toga AW, Mazziotta JC (2006): Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. Neuroimage 29:185–202.

Cocosco CA, Kollokian V, Kwan RK, Evans AC (1997): BrainWeb: Online interface to a 3D MRI simulated brain database. Neuroimage 5:425.

Collins DL, Zijdenbos AP, Kollokian V, Sled J, Kabani NJ, Holmes CJ, Evans AC (1998): Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 17:463–468.

Cordato NJ, Duggins AJ, Halliday GM, Morris JG, Pantelis C (2005): Clinical deficits correlate with regional cerebral atrophy in progressive supranuclear palsy. Brain 128 (Part 6):1259–1266.

Counsell SJ, Boardman JP (2005): Differential brain growth in the infant born preterm: Current knowledge and future developments from brain imaging. Semin Fetal Neonatal Med 10:403–410.

Cuadra MB, Cammoun L, Butz T, Cuisenaire O, Thiran JP (2005): Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Trans Med Imaging 24:1548–1565.

Dale AM, Fischl B, Sereno MI (1999): Cortical surface-based analysis I: Segmentation and surface reconstruction. Neuroimage 9:179–194.

Droske M, Meyer B, Rumpf M, Schaller C (2005): An adaptive level set method for interactive segmentation of intracranial tumors. Neurol Res 27:363–370.

Duncan JS, Papademetris X, Yang J, Jackowski M, Zeng X, Staib LH (2004): Geometric strategies for neuroanatomic analysis from MRI. Neuroimage 23 (Suppl 1):34–45.

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 (2006): Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. Hum Brain Mapp 2:99–113.

Fischl B, Sereno MI, Dale AM (1999): Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9:195–207.

Fischl B, Salat D, Busa E, Albert M, Dietrich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM (2002): Whole brain segmentation. Automated labeling of neuroanatomical structures in the human brain. Neuron 33:341–355.

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat D, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM (2004): Automatically parcellating the human cerebral cortex. Cereb Cortex 14:11–22.

Fjell AM, Walhovd KB, Reinvang I, Lundervold A, Dale AM, Quinn BT, Makris N, Fischl B (2005): Age does not increase rate of forgetting over weeks: Neuroanatomical volumes and visual memory across the adult life-span. J Int Neuropsych Soc 11:2–15.

Fotenos AF, Snyder AZ, Girton LE, Morris JC, Buckner RL (2005): Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 64:1032–1039.

Frangou S, Chitins X, Williams SC (2004): Mapping IQ and gray matter density in healthy young people. Neuroimage 3:800–805.

Good CD, Scahill RI, Fox NC, Ashburner J, Friston KJ, Chan D, Crum WR, Rossor MN, Frackowiak RS (2002): Automatic differentiation of anatomical patterns in the human brain: Validation with studies of degenerative dementias. Neuroimage 17: 29–46.

Grabowski TJ, Frank RJ, Szumski NR, Brown CK, Damasio H (2000): Validation of partial tissue segmentation of single-channel magnetic resonance images of the brain. Neuroimage 12:640–656.

Grant Steen R, Mull C, McClure R, Hamer RM, Lieberman JA (2006): Brain volume in first-episode schizophrenia. Br J Psychiat 188:510–518.

Greenspan H, Ruf A, Goldberger J (2006): Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. IEEE Trans Med Imaging 25:1233–1245.

Harris G, Andreasen NC, Cizadlo T, Bailey JM, Bockholt HJ, Magnotta VA, Arndt S (1999): Improving tissue classification in MRI: A three-dimensional multispectral discriminant analysis method with automated training class selection. J Comput Assist Tomogr 23:144–154.

Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A (2006): Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33:115–126.

Henriksson KM, Wickstrom K, Maltesson N, Ericsson A, Karlsson J, Lindgren F, Astrom K, McNeil TF, Agartz I (2006): A pilot study of facial, cranial and brain MRI morphometry in men with schizophrenia. Psychiatry Res 147 (Part 2):187–195.

Honea R, Crow TJ, Passingham D, Mackay CE (2005): Regional deficits in brain volume in schizophrenia: A meta-analysis of voxel-based morphometry studies. Am J Psychiatry 162:2233–2245.

Job DE, Whalley HC, McConnell S, Glabus M, Johnstone EC, Lawrie SM (2002): Structural gray matter differences between first-episode schizophrenics and normal controls using voxel-based morphometry. Neuroimage 17:880–889.

John NM, Kabuka MR, Ibrahim MO (2003): Multivariate statistical model for 3D image segmentation with application to medical images. J Digit Imaging 16:365–377.

Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab′bagh Y, MacDonald D, Lee JM, Kim SI, Evans AC (2005): Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. Neuroimage 27:210–221.

Kovacevic N, Lobaugh NJ, Bronskill MJ, Levine B, Feinstein A, Black SE (2002): A robust method for extraction and automatic segmentation of brain images. Neuroimage 17:1087–1100.

Kwan RK, Evans AC, Pike GB (1999): MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans Med Imaging 18:1085–1097.

Lemieux L, Hammers A, Mackinnon T, Liu RSN (2003): Automatic segmentation of the brain and intracranial cerebral spinal fluid in -weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. Magn Reson Imaging 49:872–884.

Leow A, Yu CL, Lee SJ, Huang SC, Protas H, Nicolson R, Hayashi KM, Toga AW, Thompson PM (2005): Brain structural mapping using a novel hybrid implicit/explicit framework based on the level-set method. Neuroimage 24:910–927.

Lie J, Lysaker M, Tai XC (2006): A binary level set model and some applications to Mumford-Shah image segmentation. IEEE Trans Image Process 15:1171–1181.

Lüders E, Steinmetz H, Jäncke L (2002): Brain size and grey matter volume in the healthy human brain. Neuroreport 13:2371–2374.

Lundervold A, Taxt T, Ersland L, Fenstad AM (2000): Volume distribution of cerebrospinal fluid using multispectral MR imaging. Med Image Anal 4:123–136.

Magnotta VA, Friedman L, FIRST BIRN(2006): Measurement of signal-to-noise and contrast-to-noise in the fBIRN multicenter imaging study. J Digit Imaging 2:140–147.

Makris N, Kaiser J, Haselgrove C, Seidman LJ, Biederman J, Boriel D, Valera EM, Papadimitriou GM, Fischl B, Caviness VS, Kennedy DN (2006): Human cerebral cortex: A system for the integration of volume- and surface-based representations. Neuroimage 33:139–153.

Marcelis M, Suckling J, Hofman P, Woodruff P, Bullmore E, van Os J (2006): Evidence that brain tissue volumes are associated with HVA reactivity to metabolic stress in schizophrenia, Schizophr Res 86:45–53.

Meyer-Lindenberg A, Mervis CB, Sarpal D, Koch P, Steele S, Kohn P, Marenco S, Morris CA, Das S, Kippenhan S, Mattay VS, Weinberger DR, Berman KF (2005): Functional, structural, and metabolic abnormalities of the hippocampal formation in Williams syndrome. J Clin Invest 115:1888–1895.

Meyer-Lindenberg A, Buckholtz JW, Kolachana B, Hariri AR, Pezawas L, Blasi G, Wabnitz A, Honea R, Verchinski B, Callicott JH, Egan M, Mattay V, Weinberger DR (2006): Neural mechanisms of genetic risk for impulsivity and violence in humans. Proc Natl Acad Sci USA 103:6269–6274.

Moretti B, Fadili LM, Ruan S, Bloyet N, Mazoyer B (2000): Phantom-based performance evaluation: Application to brain segmentation from magnetic resonance images. Med Image Anal 4:303–316.

Nishida M, Makris N, Kennedy DN, Vangel M, Fischl B, Krishnamoorthy KS, Caviness VS, Grant PE (2006): Detailed semiautomated MRI based morphometry of the neonatal brain: Preliminary results. Neuroimage 32:1041–1049.

Pezawas L, Meyer-Lindenberg A, Drabant EM, Verchinski BA, Munoz KE, Kolachana BS, Egan MF, Mattay VS, Hariri AR, Weinberger DR (2005): 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: A genetic susceptibility mechanism for depression. Nat Neurosci 8:828–834.

Pham DL, Xu C, Prince JL (2000): Current methods in medical image segmentation. Ann Rev Biomed Eng 2:315–337.

Rehm K, Schaper K, Anderson J, Woods R, Stoltzner S, Rottenberg D (2004): Putting our heads together: A consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. Neuroimage 22:1262–1270.

Ridha BH, Barnes J, Bartlett JW, Godbolt A, Pepple T, Rossor MN, Fox NC (2006): Tracking atrophy progression in familial Alzheimer's disease: A serial MRI study. Lancet Neurol 5:828–834.

Saeed N (1998): Magnetic resonance image segmentation using pattern recognition, and applied to image registration and quantitation. NMR Biomed 11:157–167.

Sepulcre J, Sastre-Garriga J, Cercignani M, Ingle GT, Miller DH, Thompson AJ (2006): Regional gray matter atrophy in early primary progressive multiple sclerosis: A voxel-based morphometry study. Arch Neurol 8:1175–1180.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy R, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23:208–219.

Sporn AL, Greenstein DK, Gogtay N, Jeffries NO, Lenane M, Gochman P, Clasen LP, Blumenthal J, Giedd JN, Rapoport JL (2003): Progressive brain volume loss during adolescence in childhood-onset schizophrenia. Am J Psychiatry 160:2181–2189.

Szabo CA, Lancaster JL, Lee S, Xiong JH, Cook C, Mayes BN, Fox PT (2006): MR imaging volumetry of subcortical structures and cerebellar hemispheres in temporal lobe epilepsy. AJNR Am J Neuroradiol 27:2155–2160.

Taxt T, Lundervold A (1994): Multispectral analysis of the brain using magnetic resonance imaging. IEEE Trans Med Imag 13:470–481.

Toga AW, Thompson PM (2003): Temporal dynamics of brain anatomy. Ann Rev Biomed Eng 5:119–145.

Vannier MW, Butterfield RL, Jordan S, Murphy WA, Levitt RG, Gado M (1985): Multispectral analysis of magnetic resonance images. Radiology 154:221–224.

Walhovd KB, Fjell AM, Reinvang I, Lundervold A, Fischl B, Salat D, Quinn BT, Makris N, Dale AM (2005): Cortical volume and speed-of-processing are complementary in prediction of performance intelligence. Neuropsychologia 43:704–713.

Wang D, Doddrell DM (2002): MR image-based measurement of rates of change in volumes of brain structures. I. Method and validation. Magn Reson Imaging 20:27–40.

Warfield SK, Zou KH, Wells WM (2004): Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23:903–921.

Zaidi H, Ruest T, Schoenahl F, Montandon ML (2006): Comparative assessment of statistical brain MR image segmentation algorithms and their impact on partial volume correction in PET. Neuroimage 32:1591–1607.

Zhang Y, Brady M, Smith S (2001): Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. IEEE Trans Med Imaging 20:45–57.

Zijdenbos AP, Dawant BM (1994): Brain segmentation and white matter lesion detection in MR images. Crit Rev Biomed Eng 22:401–465.