Published in final edited form as:
Neuroimage. 2006 May 1; 30(4): 1187–1195.

Statistical Parametric Mapping of Brain Morphology: Sensitivity is Dramatically Increased by Using Brain-Extracted Images as Inputs

George Fein[1,*], **Bennett Landman**[1], **Hoang Tran**[1], **Jerome Barakos**[2], **Kirk Moon**[2], **Victoria Di Sclafani**[1], and **Robert Shumway**[3]

1 *Neurobehavioral Research, Inc., Corte Madera, California*

2 *Department of Radiology, California Pacific Medical Center, San Francisco, California*

3 *Department of Statistics, University of California, Davis*

Abstract

A major attraction of voxel-based morphometry (VBM) is that it allows researchers to explore large datasets with minimal human intervention. However, the validity and sensitivity of the Statistical Parametric Mapping (SPM2) approach to VBM is the subject of considerable debate. We visually inspected the SPM2 gray matter segmentations for 101 research participants and found a gross inclusion of non-brain tissue surrounding the entire brain as gray matter in five subjects, and focal areas bordering the brain in which non-brain tissue was classified as gray matter in many other subjects. We also found many areas in which the cortical grey matter was incorrectly excluded from the segmentation of the brain. The major source of these errors was the misregistration of individual brain images with the reference T1-weighted brain template. These errors could be eliminated if SPM2 operated on images from which non-brain tissues (scalp, skull, and meninges) are removed (brain-extracted images). We developed a modified SPM2 processing pipeline that used brain-extracted images as inputs to test this hypothesis. We describe the modifications to the SPM2 pipeline that allow analysis of brain-extracted inputs. Using brain-extracted inputs eliminated the non-brain matter inclusions and the cortical gray matter exclusions noted above, reducing the residual mean square errors (RMSEs, the error term of the SPM2 statistical analyses) by over thirty percent. We show how this reduction in the RMSEs profoundly affects power analyses. SPM2 analyses of brain-extracted images may require sample sizes only half as great as analyses of non-brain extracted images.

Keywords

brain segmentation; voxel-based morphometry (VBM); statistical probability mapping; SPM2

Introduction

Voxel-based morphometry (VBM) using Statistical Parametric Mapping (SPM2) is a fully automated approach to the statistical assessment of differences (usually atrophic) between groups of subjects in their magnetic resonance images (MRI) of the brain. SPM2 uses T1-weighted brain images that include the non-brain tissues of scalp, skull, and meninges as input. The goal of VBM is to separate relevant differences in brain tissue from normal anatomic variation, artifact, and noise (Ashburner and Friston, 2000;Good et al., 2001;Wright et al., 1995).

*Corresponding author George Fein, Ph.D., President and Senior Scientist, Neurobehavioral Research, Inc. 201 Tamal Vista Blvd, Corte Madera, CA 94925 Ph: (415) 927-7676 FAX: (415) 924-2903 e-mail: george@nbresearch.com

The power of SPM2 lies in its assumption that model error terms are normally distributed and that the set of residuals is spatially distributed as a Gaussian field. This enables hypothesis testing on individual voxels with correction for multiple comparisons (Salmond et al., 2002). The important advantages over whole brain or region-of-interest structural image analysis methods are *1)* the ability to localize structural differences with weak, or no, *a priori* assumptions, *2)* the computation of confidence intervals on results, and *3)* the incorporation of an automated tissue classification algorithm.

A major attraction of VBM is that it allows researchers to explore large datasets with minimal human intervention; however, its validity and sensitivity are the subject of considerable debate. Although its source code is openly available, the source code is highly complex and requires a large investment of time to understand, examine, and modify its component parts. This has slowed progress in assessing SPM2's validity and sensitivity. Many investigators use SPM2 as a 'black box', influencing its output by the choice of input images and system options, followed by a non-quantitative visual inspection of its results. Some assert that competent identification of errors by visual inspection is a hopeful assumption at best, given the *ad hoc* nature of these visual approaches (Bookstein, 2001;Crum et al., 2003).

Practical testing of the performance of SPM2 under various conditions has proven to be a daunting task. Wilke et al. (2003) screened numerous variations in SPM2 system options for the ability to detect malformations of cortical development in epilepsy patients (using neuroradiological readings of cortical dysplasias as the gold standard), and found that the specific choice of preprocessing settings (smoothing, modulation, normalization) resulted in a variety of detection sensitivities. Mehta et al. (2003) evaluated the ability of SPM2 to detect focal lesions compared to expert classification and found that SPM2 was unable to reliably identify lesions "due to the adverse influence of lesions on preprocessing steps, and to insufficient statistical power". Tisserand et al. (2002) evaluated SPM2 using manual identification and voluming of pre-defined regions as the gold standard and concluded that although voxel-based methods can provide a reasonable estimate of regional brain volume, they cannot serve as a substitute for manual volumetry. In contrast, Testa et al. (2004) reported that SPM2 detected more instances of hippocampal atrophy in Alzheimer's disease than a comparison region-of-interest based method. In summary, these studies found that for some research questions SPM2 may have greater sensitivity than region-of-interest methods, while for other questions it may have reduced sensitivity; however, none of these reports suggested that SPM2 produces false results.

From a theoretical perspective, Bookstein argued that registration and normalization errors in the neighborhood of tissue boundaries can profoundly affect the outcome of statistical analyses (Bookstein, 2001). Ashburner and Friston (2001) agreed that such errors likely arise in VBM, but maintain that these errors should not be associated with group membership (bias), but should only add to error (increasing noise, and therefore decreasing sensitivity). Recently, a number of updated and optimized methods have been introduced to minimize registration and normalization errors (often due to ventricle size and segmentation failures around the border of the brain) by allowing use of custom templates (Davis et al., 2004;Good et al., 2001).

In our use of SPM2 for VBM, we found that a major source of error was the misregistration of each individual's brain image to the reference T1-weighted brain template. We hypothesized that these errors would be diminished if SPM2 operated on images from which non-brain tissue (scalp, skull, and meninges) was removed (brain-extracted images). We developed a modified SPM2 processing pipeline that used brain-extracted images as inputs to test this hypothesis.

In this manuscript, we describe the modifications to the SPM2 pipeline that allow analysis of brain-extracted inputs. We also evaluate the effects (both qualitatively and quantitatively) of

these changes through: *1)* a qualitative visual inspection of each subject's gray matter segmentation, and *2)* a quantitative analysis of residual mean square errors, which are the error terms of the SPM2 General Linear Model statistical analyses (therefore directly associated with SPM2's sensitivity).

Methods

Subjects

This study examines recent MRI data in abstinent alcoholics and controls. Each MRI was read by a neuroradiologist. Scans were excluded for abnormalities other than white matter signal hyperintensities. The images were normal observations that did not violate the SPM2 assumptions (e.g., imaging artifact). The institutional review board approved all procedures, and written consent was obtained from all individuals prior to study.

Image Acquisition

All studies were performed on a 1.5T GE Signa Infinity with the LX platform (GE Medical Systems, Waukesha, WI) located at the Pacific Campus of the California Pacific Medical Center. The imaging protocol included a transaxial T1-weighted Spoiled Gradient image (TR/TE/NEX = 35/5/1; 0.859 × 0.859 mm^2 in-plane resolution; contiguous 1.3 mm thick slices).

Optimized VBM Analysis

Optimized VBM was implemented in the framework of SPM2 (Good et al., 2001), testing for gray matter differences between groups and included age, years of education, and cranium size as covariates. The difference in tissue volumes associated with the normal variation in cranium size was removed using the inverse of the FSL v-scaling parameter (Smith et al., 2002), which we have shown previously is an excellent surrogate variable for the size of the intracranial vault (Fein et al., 2004). The analysis included all voxels in the brain that segmented as more than 15% gray matter.

Examination of VBM Performance

We visually inspected the SPM2 gray matter segmentations (prior to scaling and smoothing) for all 101 research participants (see Figure 1). A gross inclusion of non-brain tissue surrounding the entire brain as gray matter was apparent in five subjects (gross non-brain matter inclusions – gross NBMIs). In addition, for many other individuals, there were focal areas bordering the brain in which non-brain tissue was classified as gray matter (subtle NBMIs). Figure 2 shows the sagittal T1-weighted image and the gray matter segmentation for three participants, illustrating gross NBMIs, subtle NBMIs and zero NBMIs. The relatively high amount of scalp fat (bright signal) around the boundary of the brain in the T1 images of the five individuals with gross NBMIs was striking. Based on this observation, we hypothesized that the body mass index (BMI) was elevated in individuals with gross NBMIs. Table 1 presents the BMI and other demographics for the five individuals with gross NBMIs.

We performed a Monte-Carlo simulation to evaluate the hypothesis of elevated BMIs in individuals with gross NBMIs. We constructed 10,000 random combinations of 3 females and 2 males from the 100 participants for whom we had BMI values (the five segmentation failures occurred in 3 females and 2 males). The sum of the BMIs for the five participants with gross NBMIs was greater than the sum of 9366 of the 10,000 quintuples, yielding a probability of < 0.064 that the gross NBMIs were unrelated to a participant's BMI.

Since identification of non-brain tissue as gray matter appeared to be the primary error in the segmentations, we hypothesized that misalignment of individuals' brains with the brain in the MNI152 template was the culprit. Figures 3a, 3b, and 3c illustrate: *1)* gross misalignment of

the MNI152 template brain with the participant's brain for a gross NBMI individual, *2)* focal misalignments of the template brain for a subtle NBMI subject, and *3)* the correct alignment of the template brain in a zero NBMI subject. Figure 3 illustrates that misalignment also can result in exclusion of cortical gray matter (see the subtle NBMI subject), not just in inclusion of non-brain tissue.

We tested the hypothesis that the presence of non-brain tissue in the input images and MNI template causes brain misalignment by modifying the SPM2 processing pipeline to use brain-extracted images and a brain-extracted MNI152 T1-weighted reference template. Non-brain tissue was removed from each subject's MRI using FSL's Brain Extraction Tool (Smith, 2002) (with default settings) followed by manual removal of any additional non-brain tissue missed by BET using an in-house custom written plugin to Image J (Rasband, 2002). We also used the brain-extracted MNI152 template provided by FSL (rather than the MNI152 template that is provided by SPM2, which includes the non-brain tissues of scalp, skull, and meninges). The FSL template was smoothed using a 12 mm FWHM Gaussian kernel to match the smoothing of the SPM2 template before inserting it into the SPM2 pipeline. Figures 4a, 4b, and 4c show that the outer boundaries of the SPM2 segmentations align with the outer boundary of the brain when SPM2 uses brain-extracted inputs for the subjects in Figures 3a–c. Figure 5 presents, for all 101 participants, the difference image subtracting the SPM2 segmentation computed with brain-extracted inputs from those computed with non-brain extracted inputs, displaying only the positive values as white. This figure illustrates the magnitude and prevalence of the effect of non-brain matter inclusions on the gray matter segmentations. Figure 6 presents the difference images subtracting the SPM2 segmentations computed with non-brain extracted inputs from those computed with brain-extracted inputs, also only displaying the positive values as white. This figure illustrates the magnitude and prevalence of the effect of incorrect exclusion of gray matter on the gray matter segmentations (as displayed in Figure 3b, above). In the analysis of non-brain extracted input images, the incorrect exclusion of gray matter appears to be as big a problem as are non-brain matter inclusions.

Residual Mean Square Errors (RMSEs)

We denote the theoretical average RMSEs for each SPM2 implementation by σ_{i1}^2 and σ_{i2}^2, and the measured RMSEs by s_{i1}^2 and s_{i2}^2. Figures 7 and 8 present the measured RMSEs for SPM2 using brain-extracted inputs versus non-brain extracted inputs. The histograms of the RMSEs across the approximately 1.5 million observations (one for each voxel) are also presented in these figures. These figures illustrate dramatically reduced RMSEs result from using brain-extracted inputs.

The mean RMSEs are $\bar{s}_1^2 = .2098$ and $\bar{s}_2^2 = .4708$, with variances $v_1 = .0132$ and $v_2 = .0758$. The average RMSE using brain-extracted inputs is less than half the average RMSE using non-brain extracted inputs, and the variance using brain-extracted inputs is about one sixth the variance using non-brain extracted images. SPM2 test statistics for main effect and interactions will have standard deviations proportional to 0.4580 and 0.6861 (√.2098 and √.4708). If brain-extracted image inputs are used, SPM2 can detect effect sizes that are about one third smaller than those that can be detected using non-brain extracted inputs (0.4580/0.6861)=0.6675.

Given that the images have about 100 independent observations, and letting $n = \min(n_1, n_2) \approx 100$,

$$z = \frac{(\bar{s}_1^2 - \bar{s}_2^2)}{\sqrt{\frac{v_1}{n_1} + \frac{v_2}{n_2}}}$$

will have an approximately normal distribution for reasonably large n. In this case, we obtain $z = -5.96$ which has a P-value of $\ll 0.001$; a very strong rejection of the hypothesis of equal average RMSEs. In fact, for any $n > 16$, the 0.001 critical value for $z = -3.09$ is dramatically exceeded. With $n \approx 100$, the α will be orders of magnitude smaller than 0.001.

Discussion

SPM2 is an implementation of VBM that uses T1-weighted brain images (that include the scalp, skull and meninges) as inputs, and incorporates a morphological clean-up step to remove 'non-brain' tissue. We found that: *1)* SPM2 does a poor job at the removal of non-brain tissue, 2) poor alignment of individual images with the MNI template also results in incorrect exclusion of cortical gray matter, and 3*)* that both of these effects negatively impact the sensitivity of the method to detect experimental effects. Registration of each individual's MRI to a template (the MNI152 is the SPM2 default template) is the first step in the SPM2 processing pipeline. In the work reported above, we show that errors in this initial registration occur in SPM2 and negatively impact results. Fortunately, a relatively simple modification to the SPM2 pipeline can fix this misregistration problem.

Examining the SPM2 gray matter segmentation results, we observed gross segmentation errors in a number of subjects. These errors occurred primarily at the outer boundaries of the brain, where non-brain was included in the segmentation of the brain and cortical gray matter was incorrectly excluded from the segmentation of the brain. SPM2's morphological clean-up function inadequately addressed these problems. Our solution to this problem was to modify SPM2 to process brain-extracted MRIs, and to register those images to a brain-extracted template. This solution is conceptually very simple, but does involve significant work. There was nothing special (abnormal or 'below par') about our subjects' MRIs. The MRIs were clinically normal, except for the presence of white matter signal hyperintensities. However, sample characteristics may magnify the SPM2 registration errors described above. We noted that MRIs with failed segmentations tended to have an abundance of high signal scalp fat. We also found a strong statistical trend for those individuals to have a higher BMI than our other research participants. It may be that any condition or disease that affects the fat signal from the scalp (obesity, anorexia, etc.) may impact the sensitivity SPM2 results.

Senjem et al (2005) recently compared a number of different methodological implementations of SPM2 in the analysis of morphological changes in Alzheimer's disease. One of the methods he presented did include removal of non-brain tissue from the image inputs (brain extraction). However, their publication is not directly comparable to this manuscript. While Senjem helped identify important questions about the SPM2 approach to VBM, they did not: *1)* assess sensitivity or validity in a quantitative manner *2)* incorporate brain extraction prior to registration and alignment, *3)* examine segmentation results for individual subjects.

The increased sensitivity that derives from the analysis of brain-extracted images has dramatic effects on experimental power. For example, if one wanted to replicate a finding of a mean difference of d=0.8 from a non-brain extracted analysis, this would translate into an effect size of d=1.19 for the analysis of brain-extracted images. Power of 0.80 for such a replication study would require samples of 26 subjects per group for non-brain extracted images (d=0.8) and about 12 subjects per group for brain-extracted images (d=1.19). Conversely, an effect size of d=0.8 from a study with brain-extracted images (requiring the same 26 subjects per group)

would translate to an effect size of d=0.54 for non-brain extracted images (now requiring 54 subjects per group for power of 0.80). Thus, our best estimate is that sample sizes half as large are required to detect the same size effects for brain-extracted vs. non-brain extracted inputs.

In sum, complete initial removal of all non-brain tissue from brain MRIs results in more sensitive analyses of brain morphology using SPM2. We note that we have not even begun to address the controversy in the literature regarding SPM2 that deals with its use of prior probability templates in its Bayesian approach to tissue segmentation. First, the prior probability template must be appropriate for the population from which research subjects are drawn. For example, it may not be appropriate to use MNI152 tissue probability templates for samples outside of the age range from which the MNI152 sample was drawn. Second, the prior probability template used may better match tissue probabilities in one research group than another. This is likely to be the case in comparing an Alzheimer Disease sample to age comparable normal controls. The AD sample is likely to have much larger ventricles and cortical atrophy. No single template would be equally appropriate for both groups; however, each group's MRIs must be processed identically to avoid the introduction of method induced bias. One might consider replacing the Bayesian segmentation in SPM2 with a segmentation approach that does not require a prior probability template. In the manuscript presented above we provide a framework for the examination of these problems and their potential solutions.

Acknowledgements

This work was supported by Grants AA11311 (GF) and AA13659 (GF), both from the National Institute of Alcoholism and Alcohol Abuse.

References

Ashburner J, Friston KJ. Voxel-based morphometry--the methods. Neuroimage 2000;11:805–21. [PubMed: 10860804]

Ashburner J, Friston KJ. Why voxel-based morphometry should be used. Neuroimage 2001;14:1238–43. [PubMed: 11707080]

Bookstein FL. Voxel-based morphometry" should not be used with imperfectly registered images. Neuroimage 2001;14:1454–62. [PubMed: 11707101]

Crum WR, Griffin LD, Hill DL, Hawkes DJ. Zen and the art of medical image registration: correspondence, homology, and quality. Neuroimage 2003;20:1425–37. [PubMed: 14642457]

Davis, B.; Lorenzen, P.; Joshi, S. Large deformation minimum mean squared error template estimation for computational anatomy. Paper presented at the Proceedings of the IEEE Symposium on Biomedical Imaging; Arlington, VA. 2004; 2004.

Fein G, Di Sclafani V, Taylor C, Moon K, Barakos J, Tran H, Landman B, Shumway R. Controlling for premorbid brain size in imaging studies: T1-derived cranium scaling factor vs. T2-derived intracranial vault volume. Psychiatry Res 2004;131:169–76. [PubMed: 15313523]

Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 2001;14:21–36. [PubMed: 11525331]

Mehta S, Grabowski TJ, Trivedi Y, Damasio H. Evaluation of voxel-based morphometry for focal lesion detection in individuals. Neuroimage 2003;20:1438–54. [PubMed: 14642458]

Rasband, W. Image J [PC program]. National Institute of Health; USA: 2002.

Salmond CH, Ashburner J, Vargha-Khadem F, Connelly A, Gadian DG, Friston KJ. Distributional assumptions in voxel-based morphometry. Neuroimage 2002;17:1027–30. [PubMed: 12377176]

Senjem ML, Gunter JL, Shiung MM, Petersen RC, Jack CR Jr. Comparison of different methodological implementations of voxel-based morphometry in neurodegenerative disease. Neuroimage 2005;26:600–8. [PubMed: 15907317]

Smith SM. Fast robust automated brain extraction. Human Brain Mapping 2002;17:143–155. [PubMed: 12391568]

Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, De Stefano N. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 2002;17:479–89. [PubMed: 12482100]

Testa C, Laakso MP, Sabattoli F, Rossi R, Beltramello A, Soininen H, Frisoni GB. A comparison between the accuracy of voxel-based morphometry and hippocampal volumetry in Alzheimer's disease. J Magn Reson Imaging 2004;19:274–82. [PubMed: 14994294]

Tisserand DJ, Pruessner JC, Sanz Arigita EJ, van Boxtel MP, Evans AC, Jolles J, Uylings HB. Regional frontal cortical volumes decrease differentially in aging: an MRI study to compare volumetric approaches and voxel-based morphometry. Neuroimage 2002;17:657–69. [PubMed: 12377141]

Wilke M, Kassubek J, Ziyeh S, Schulze-Bonhage A, Huppertz HJ. Automated detection of gray matter malformations using optimized voxel-based morphometry: a systematic approach. Neuroimage 2003;20:330–43. [PubMed: 14527593]

Wright IC, McGuire PK, Poline JB, Travere JM, Murray RM, Frith CD, Frackowiak RS, Friston KJ. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 1995;2:244–52. [PubMed: 9343609]

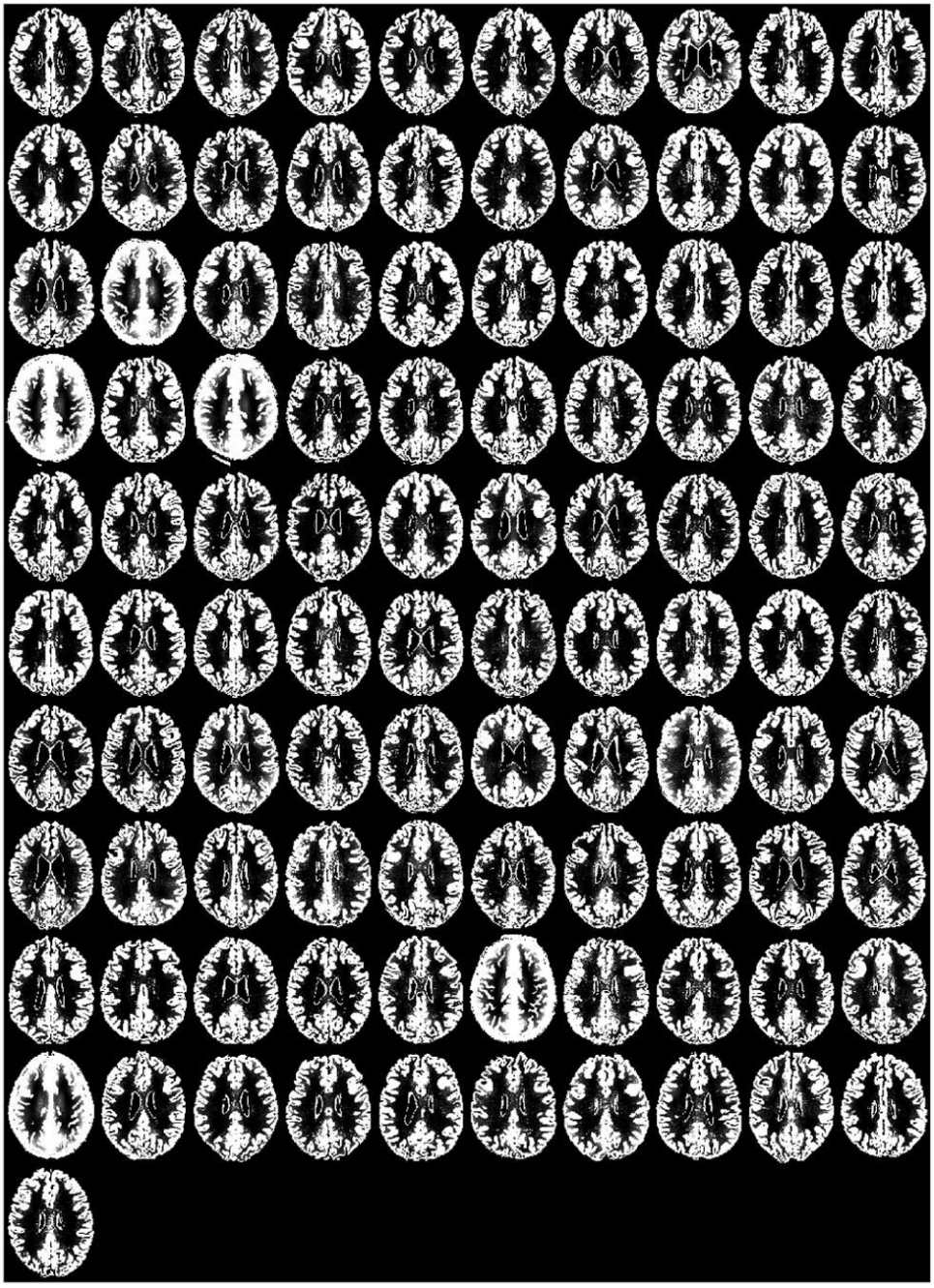


Figure 1.
Midaxial slice of the gray matter SPM2 segmentation for each of the 101 research subjects. Subjects in row 3, column 2; row 4, column 1 and 3; row 9, column 6; and row 10, column 1 stand out as having poor segmentations.

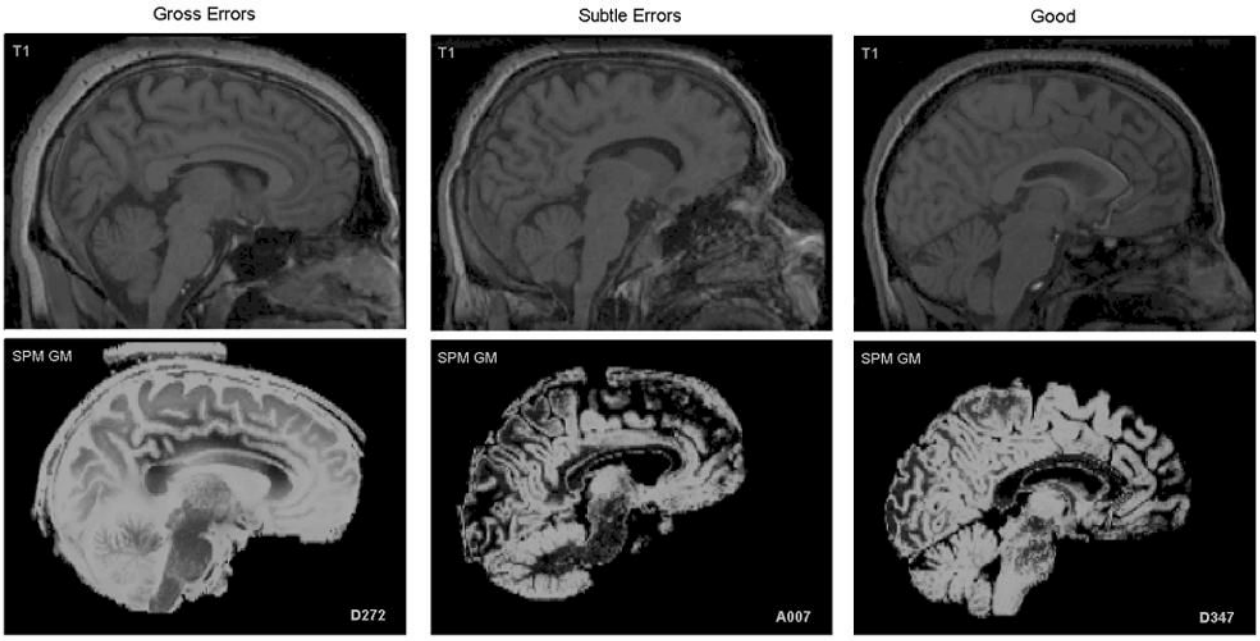
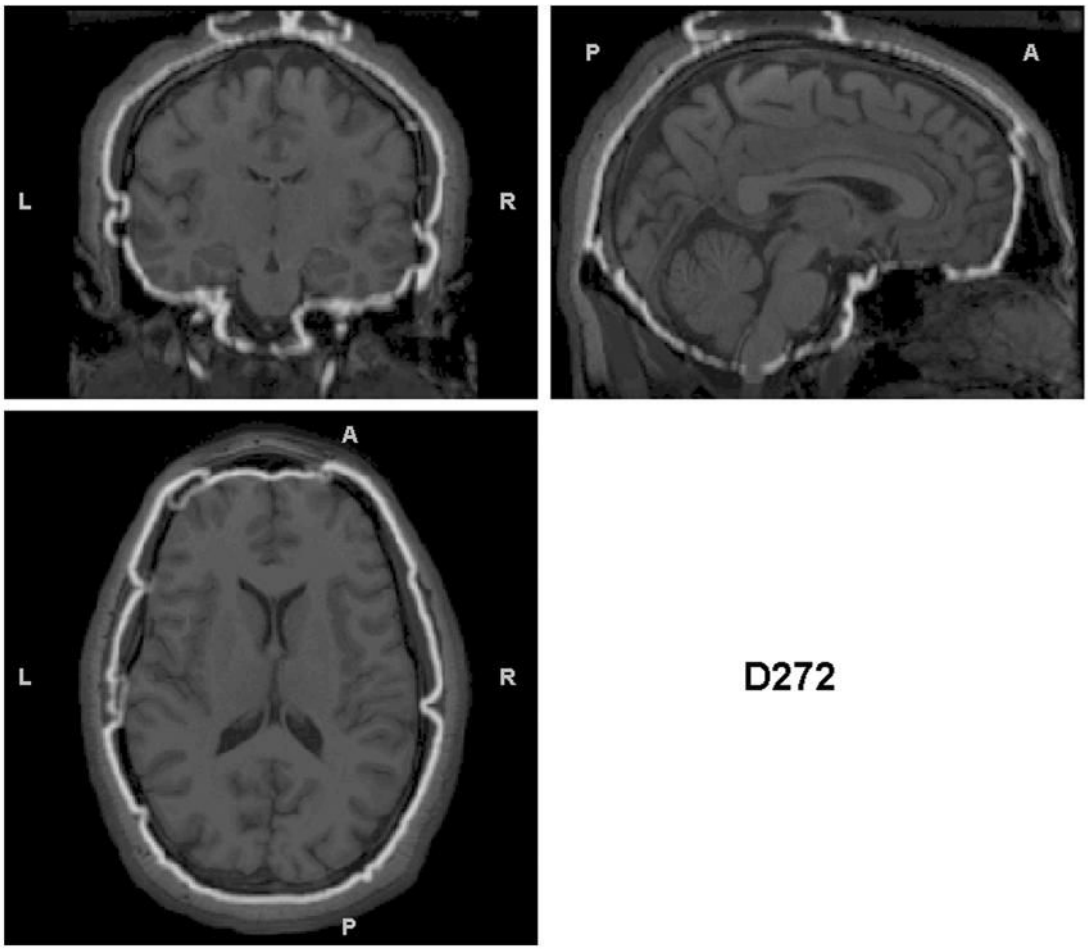
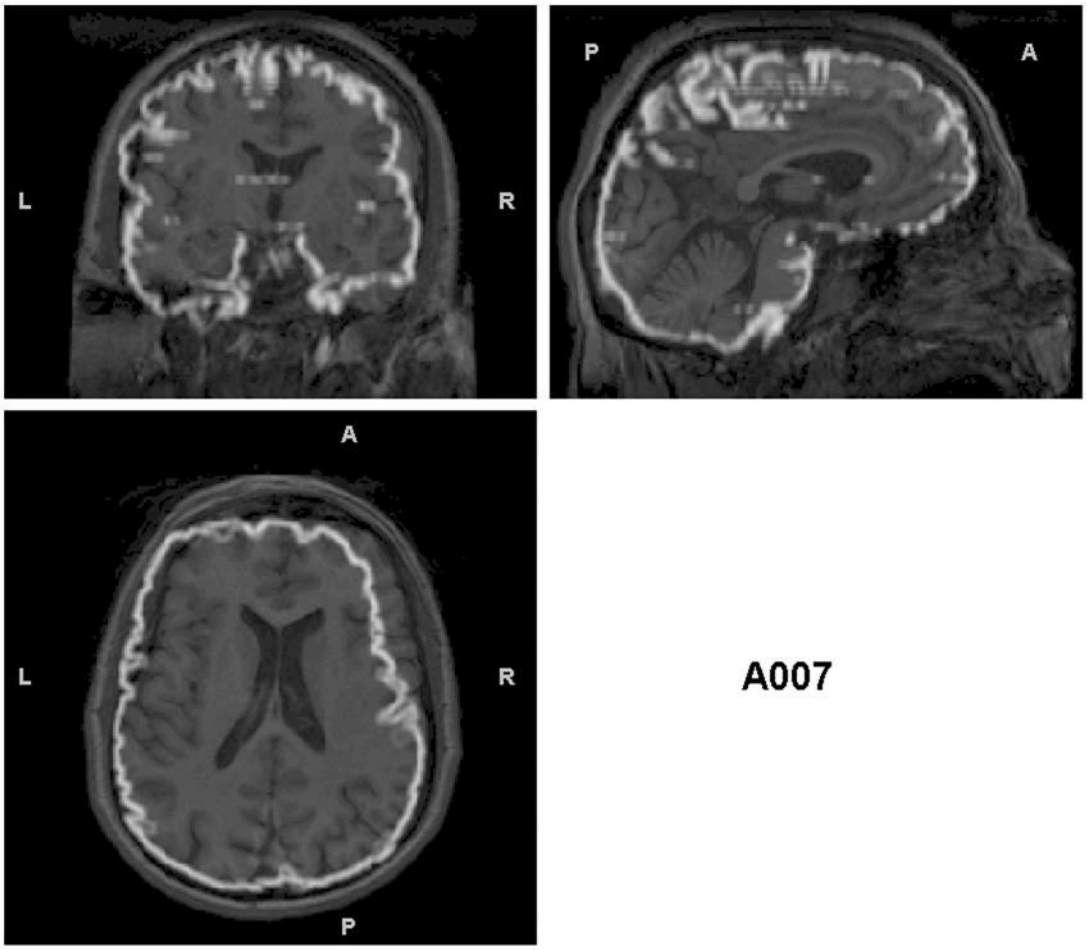


Figure 2.
Midsagittal slice of 3 subjects illustrating various degrees of segmentation errors (gross errors – left, subtle errors – middle, good segmentation – right).





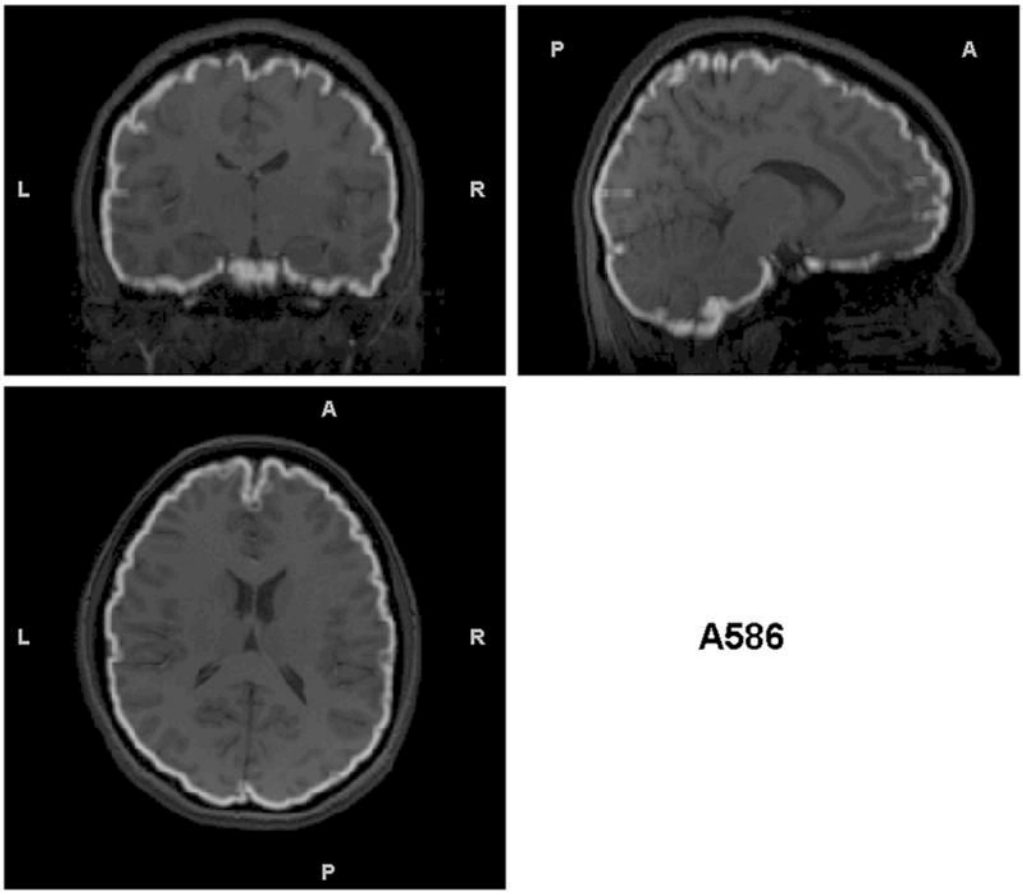
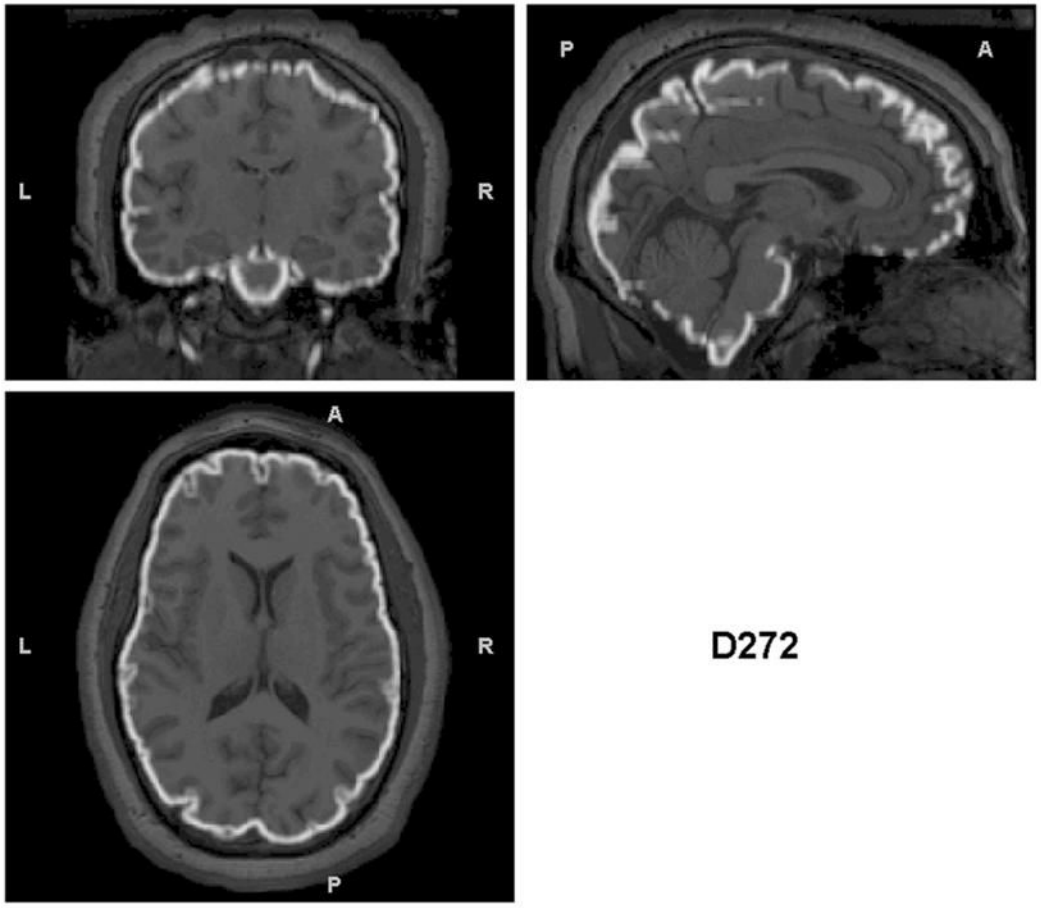
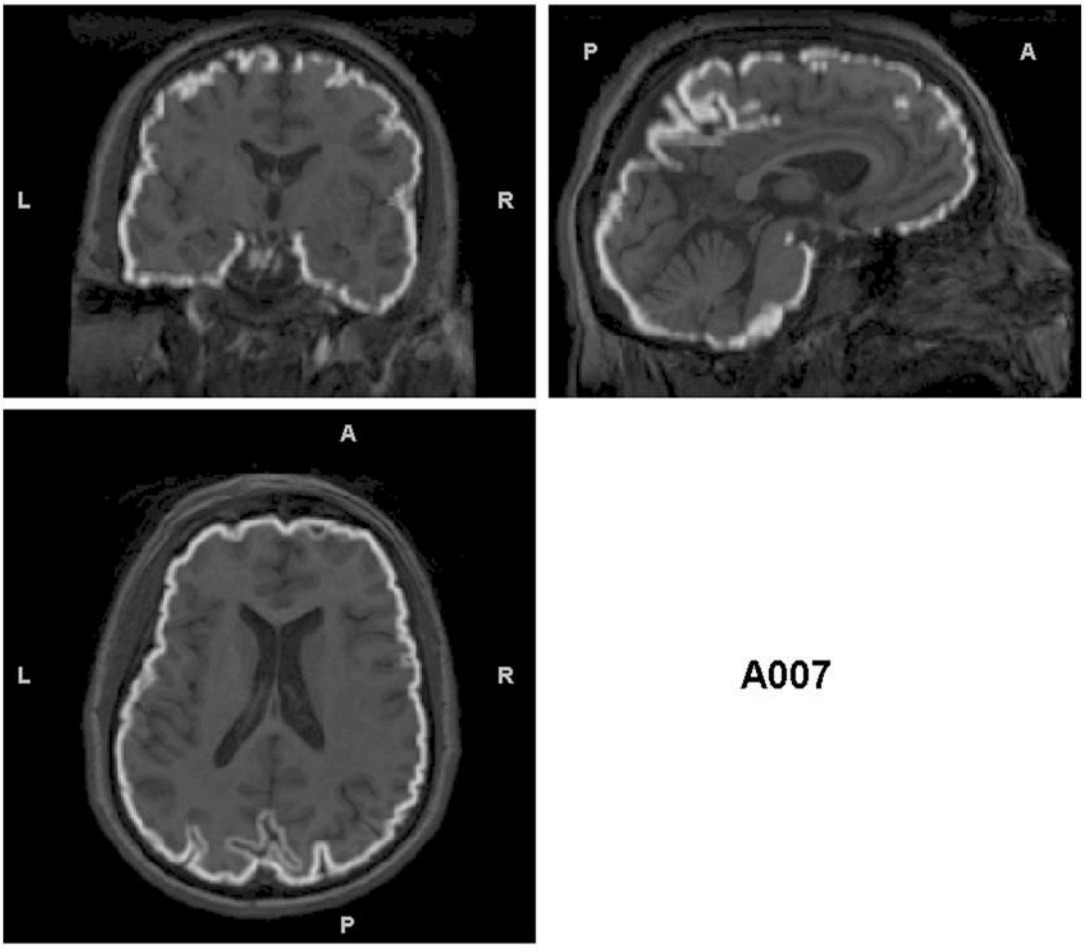


Figure 3.
Triplanar view of the outer boundary of the GM segmentation (outlined in white) using the non-brain-extracted T1s as input for the subjects presented in Figure 2. The GM segmentation for (a) is outside the border of the brain and in many cases within the skull. The segmentation in (b) has less (but some) non-brain tissue included in the outline of the GM segmentation. However, this example shows another problem that is not obvious without examining these boundaries – that there are brain regions excluded from the GM segmentation (visible in the sagittal and coronal views in the anterior right). (c) the GM segmentation is well aligned with the brain boundary, with no problems indicated.





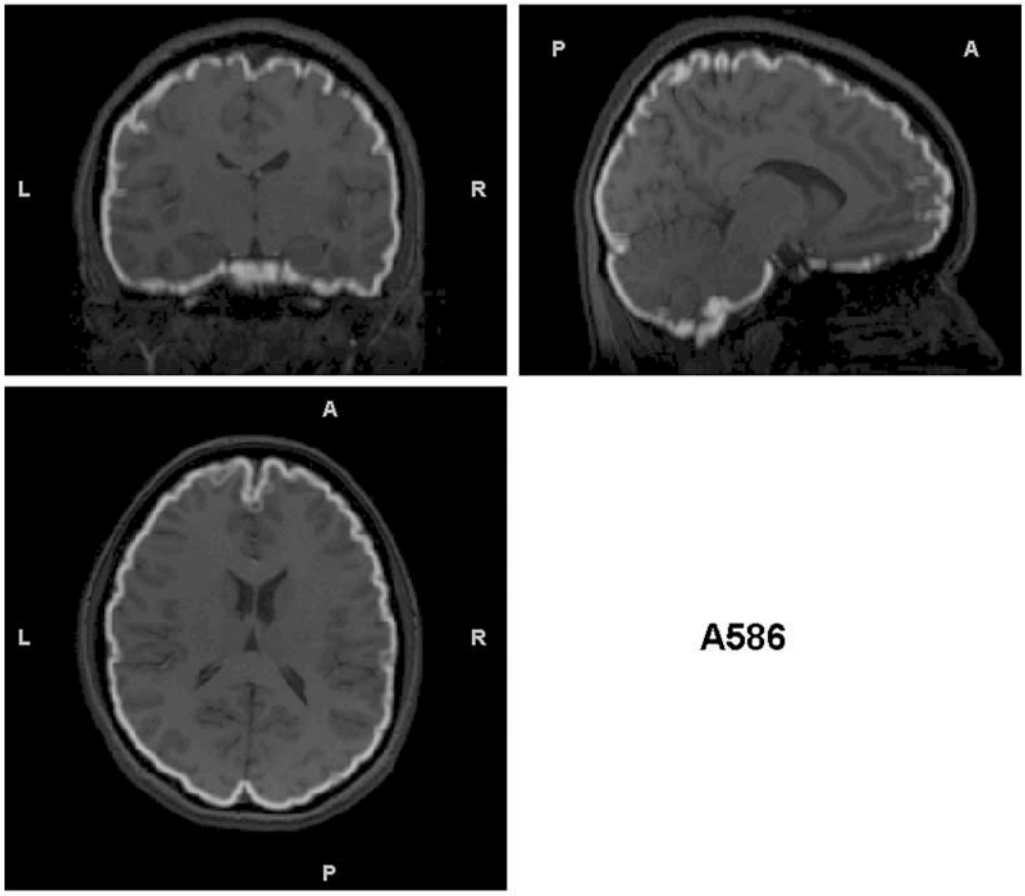


Figure 4.
Triplanar view of the outer boundary of the GM segmentation (outlined in white) using the brain-extracted T1s as input for the same subjects as in Figures 2 and 3. The GM segmentations for all subjects are well aligned with the brain boundary.

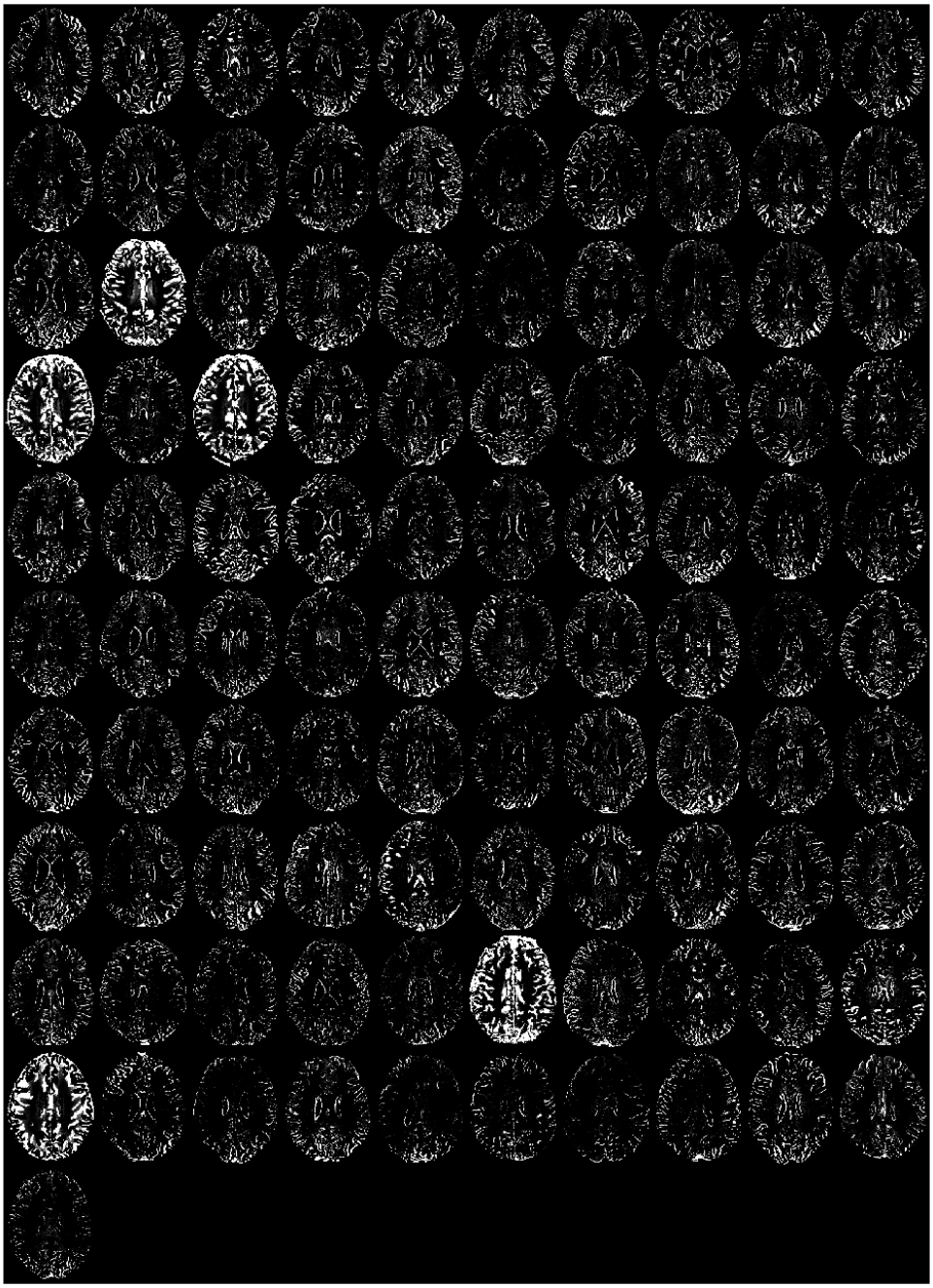


Figure 5.
Midaxial slice (for each of the 101 research subjects) of the difference in the gray matter SPM2 segmentations computed by subtracting the segmentation using brain-extracted inputs from that using non-brain extracted inputs. Bright areas denote non-brain matter inclusions in the segmentation using non-brain extracted inputs.

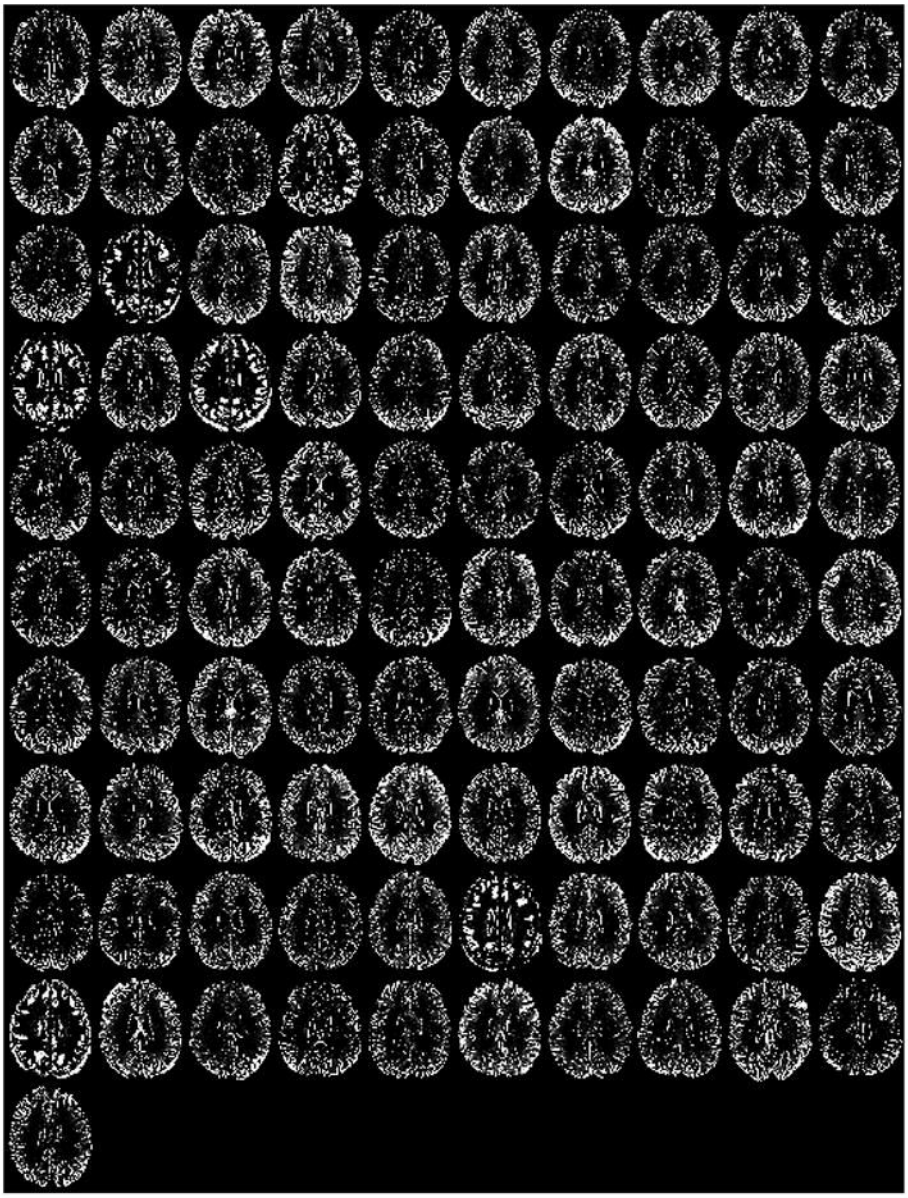


Figure 6.
Midaxial slice (for each of the 101 research subjects) of the difference in the gray matter SPM2 segmentations computed by subtracting the segmentation using non-brain extracted inputs from that of using brain-extracted inputs. Bright areas denote areas where gray matter was incorrectly excluded from the segmentation using non-brain extracted inputs.

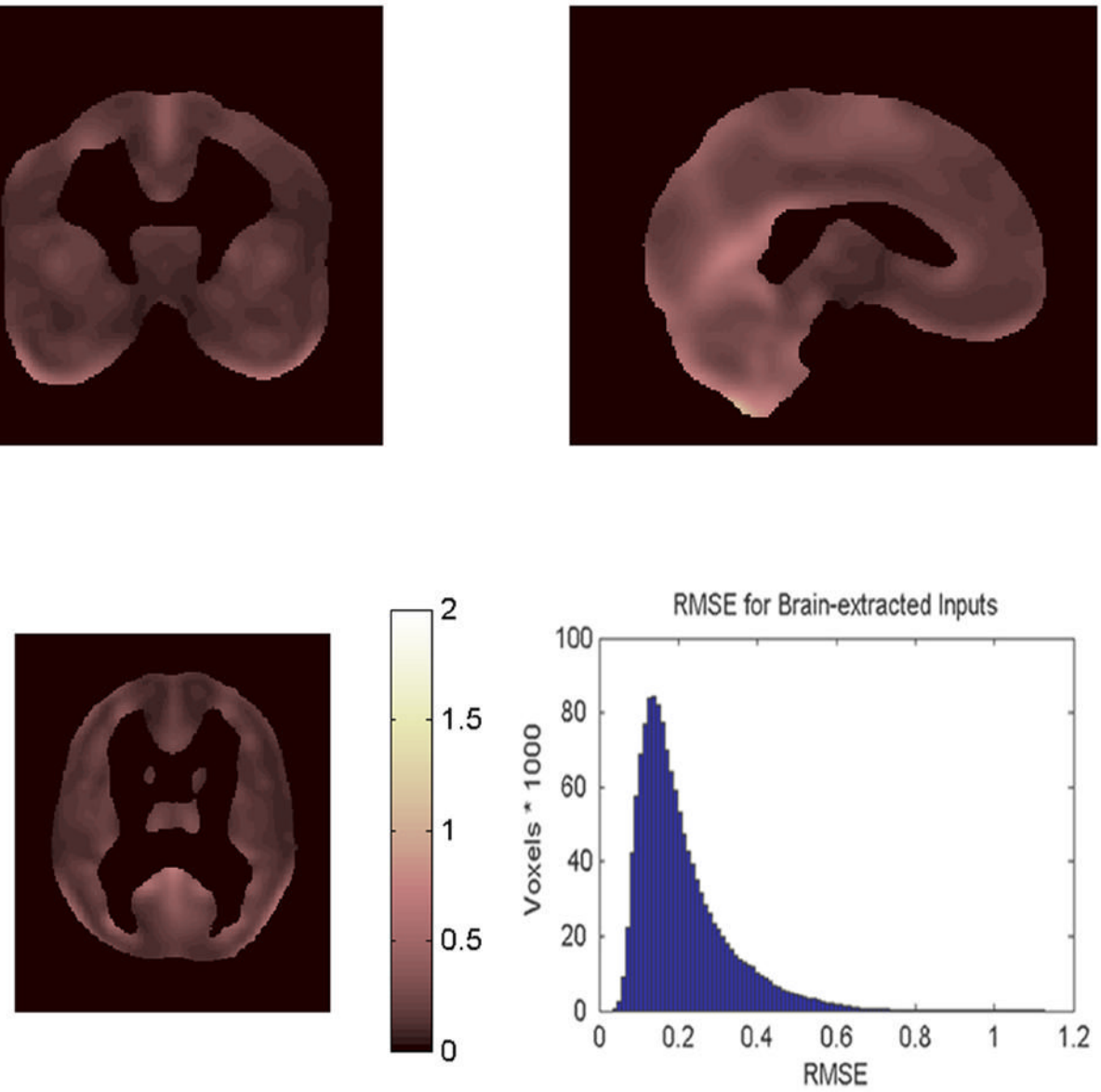


Figure 7.
Triplanar view with histogram of the Residual Mean Square Error (RMSE) for the 101 subjects using brain-extracted inputs.

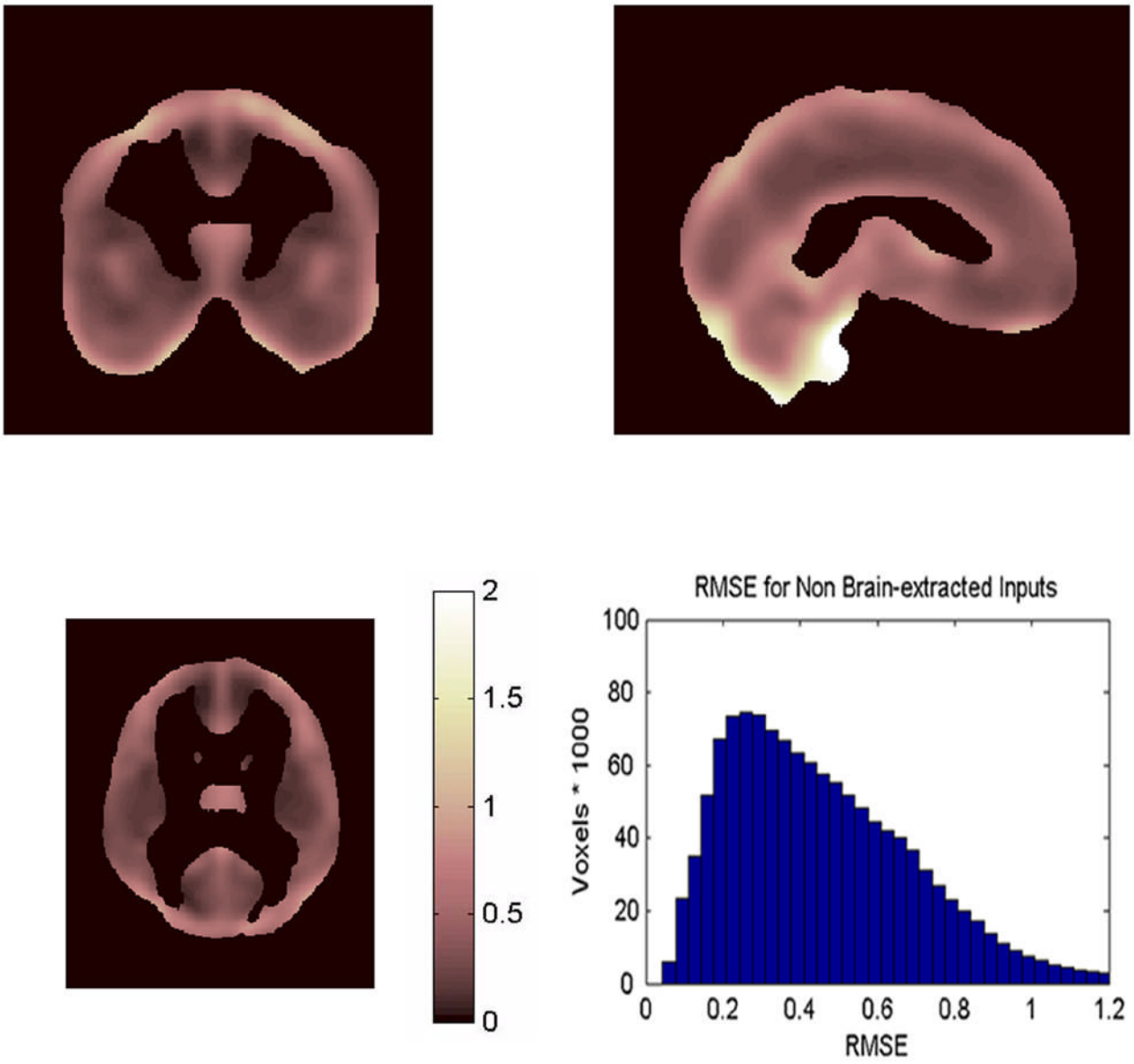


Figure 8.
Triplanar view with histogram of the Residual Mean Square Error (RMSE) for the 101 subjects using non-brain extracted inputs. Note the RMSE values are larger than for brain-extracted inputs (Figure 7), particularly near the surface boundaries.

Table 1

Demographics of Subjects with Failed Segmentations

[[]]	[[Subjects|]]				
[[Variables|]]	**[[A253|]]**	**[[A403|]]**	**[[A432|]]**	**[[A987|]]**	**[[D272|]]**
[[Group|]]	[[Abstinent|]]	[[Abstinent|]]	[[Abstinent|]]	[[Control|]]	[[Control|]]
[[Sex|]]	[[Female|]]	[[Female|]]	[[Male|]]	[[Female|]]	[[Male|]]
[[Age (years)|]]	[[51|]]	[[53|]]	[[54|]]	[[48|]]	[[35|]]
[[BMI|]] *	[[27|]]	[[27|]]	[[29|]]	[[30|]]	[[26|]]
[[Ethnicity|]]	[[C|]]	[[C|]]	[[C|]]	[[AA|]]	[[AA|]]

* The BMI (mean±S.D.) by group and sex are: 23.8±3.6 abstinent females, 27.4±3.3 abstinent males, 24.6±4.5 control females, and 24.6±3.4 control males.