Issues with threshold masking in Voxel Based Morphometry of atrophied brains.

Gerard R. Ridgway[a], Rohani Omar[b], Sébastien Ourselin[a], Derek L.G. Hill[a], Jason D. Warren[b], and Nick C. Fox[b*]

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

[b] Dementia Research Centre (DRC), Institute of Neurology, University College London, Queen Square, London WC1N 3BG, UK

Keywords: Computational anatomy; voxel based morphometry; Alzheimer's disease; threshold masking; analysis mask.

Running Title: Issues with masking in VBM of atrophy.

* **Corresponding author:**
Nick C. Fox
Dementia Research Centre
Institute of Neurology
University College London
Queen Square, London WC1N 3BG
United Kingdom

Telephone: +44 20 7829 8773
Fax: +44 20 7676 2066
E-mail: nfox@dementia.ion.ucl.ac.uk

Abstract

There is great interest in using automatic computational neuroanatomy tools to study ageing and neurodegenerative disease. Voxel-Based Morphometry (VBM) is one of the most widely used of such techniques. VBM performs voxel-wise statistical analysis of smoothed spatially normalised segmented Magnetic Resonance Images. There are several reasons why the analysis should include only voxels within a certain mask. We show that one of the most commonly used strategies for defining this mask runs a major risk of excluding from the analysis precisely those voxels where the subjects' brains were most vulnerable to atrophy. We investigate the issues related to mask construction, and recommend the use of alternative strategies which greatly decrease this danger of false negatives.

Introduction

In essence, Voxel-Based Morphometry (VBM) (Ashburner and Friston, 2000) involves voxel-wise statistical analysis of data derived from structural Magnetic Resonance (MR) brain images of multiple subjects. The images analysed are obtained through tissue segmentation, spatial normalisation, and spatial smoothing. Statistical analysis employs a mass-univariate parametric or non-parametric general linear model at each voxel. More precisely, the calculations are performed at each voxel *within some mask*. There are several reasons why masking is necessary, mostly related to the multiple comparison problem. Family-wise error (FWE) correction using random field theory (RFT) is generally more powerful for smaller analysis regions (this is commented on further in the discussion), and perhaps more importantly, masking is necessary for successful estimation of the smoothness of the residuals (John Ashburner, personal communication), which is a key part of the RFT correction procedure (Kiebel et al., 1999). If non-parametric permutation methods (Nichols and Holmes, 2002) are employed for FWE correction, the effect of the analysis region on computational complexity may also be important (Belmonte and Yurgelun-Todd, 2001). Correction of the false-discovery rate (Genovese et al., 2002) also depends on masking, since non-brain voxels could otherwise skew the distribution of p-values on which it is based. Furthermore, masking can also partially alleviate a problem of implausible false positives occurring outside the brain due to the very low variance in voxels with consistently low smoothed tissue density — the extreme limit of the phenomenon described by Reimold et al. (2006). Finally, while not specifically considered here, multivariate machine learning, classification or decoding approaches (Lao et al., 2004; Vemuri et al., 2008; Friston et al., 2008) can also benefit from

masking as an initial feature selection or dimensionality reduction step.

Having emphasised above that smaller masks generally lead to higher sensitivity and clarified interpretation, it is important to recognise the obvious risk that overly restrictive masks will lead to false negatives, as potentially interesting voxels are excluded from the statistical analysis. In this paper, we argue that there is a particular danger of false negatives arising in VBM studies of pathological brains when computing the analysis mask using a commonly employed approach with settings that appear reasonable a priori. This approach is used by the popular Statistical Parametric Mapping (SPM) software (`http://www.fil.ion.ucl.ac.uk/spm/`). We recommend the use of different mask-generation strategies, which we show to reduce this danger. In a three-part experiment using SPM, we (a) use simulated data to investigate properties of preprocessing relevant to masking; (b) explore the behaviour of standard and more novel methods of masking, considering variable patient group composition; and (c) test the practical importance of our recommendation on a particular example of a real VBM study. We propose two main masking options: one is a fully objective parameter-free algorithm, which we hope will find wide-spread applicability; the other allows expert knowledge to be exercised in cases where the automatic strategy is found to be unsuitable.

Methods

Masking strategies

The SPM software commonly used for VBM studies offers several alternatives to specify the mask for statistical analysis. If available, a precomputed mask can be explicitly requested, or the analysis mask can be automatically derived by excluding voxels in which any of the images have intensity values below a certain threshold. This threshold can be specified as an absolute value, constant for all the images, or as a relative fraction of each image's 'global' value. The global value can itself either be precomputed or can be automatically calculated as the mean of those voxel intensities which are above one eighth of the mean of all voxels. This arbitrary heuristic aims to determine an average that is not biased by the presence of potentially variable amounts of non-brain background in the field of view; it is explored below.

In VBM studies where fairly pronounced atrophy is expected, such as those of Alzheimer's Disease (Karas et al., 2003) or Semantic Dementia (Mummery et al., 2000), it is probable that some patients will have particularly low grey matter (GM) density in their most severely affected regions. It seems undesirable to exclude such regions from the statistical analysis; since this is likely to occur with SPM's threshold masking, which effectively takes the intersection of all subjects' supra-threshold voxels, we argue that a different strategy should be used for the creation of a mask (which can then be specified as an explicit mask in SPM or other software packages). We propose one such strategy here, based on the principle of replacing the criteria that all subjects should have voxel intensity above the threshold, with the relaxed requirement that some specified fraction of the subjects exhibit supra-threshold

voxel values within the mask. In other words, voxels are included if there is a consensus among some percentage of the subjects that they are above threshold. Vemuri et al. (2008) used this approach in their image classification work, with a consensus of 50% and a threshold of 0.1. SPM's method is a special case of this, where the consensus fraction is 100%.

An alternative masking strategy is to threshold the mean of all subjects' segmentations; this might be expected to be similar to using a consensus of 50%, though it is not equivalent. Here, we propose a novel idea to objectively select a threshold for the average image which optimises an intuitively reasonable objective function. Based on the observation that the average image appears to have qualitatively high probabilities over a visually appealing region, it might be expected that a good threshold T would result in the binarised mask $M = A > T$ remaining highly correlated with the unthresholded original average image A. We therefore determine an 'optimal' mask $M^* = A > T^*$ by finding the threshold that maximises this criterion:[1]

$$T^* = \arg\max_{\mathrm{T}} \rho(\mathrm{A}, \mathrm{A} > \mathrm{T}), \tag{1}$$

where $\rho(x, y)$ is the sample Pearson correlation (over pairs of voxels) between two images x and y. This strategy has neither a tunable threshold nor a specified consensus fraction, making it truly operator-independent.

[1]This simple maximisation problem can be solved with standard routines, for example using MATLAB's fminbnd to search for the best threshold between 0 and 1.

Quantitative results using simulated images

The first experiment uses artificially generated MR images from the BrainWeb project (Cocosco et al., 1997; Aubert-Broche et al., 2006),[2] derived from real MRIs of normal healthy subjects. These images have known underlying tissue segmentation models, allowing quantitative evaluation of segmentation accuracy for each simulated subject. Given some quantitative metric, we can therefore determine the optimal level at which a probabilistic segmentation must be binarised. By considering the simple Jaccard Similarity coefficient (Crum et al., 2006) between the binary (maximum probability) model of grey matter B and the estimated probabilistic segmentation S after binarisation at a particular threshold T, the optimal threshold may then be found as

$$T^* = \arg\max_{\mathrm{T}} \mathrm{J}(\mathrm{B}, \mathrm{S} > \mathrm{T}),$$
$$J(x, y) = \frac{|x \cap y|}{|x \cup y|},$$

where $|x|$ denotes the number of non-zero voxels in an image x, and the intersection and union involve voxel-wise Boolean operations. We investigate the variation in the optimal threshold with relation to preprocessing, visually and in terms of its proportionality to the global or total signal. SPM's estimated global averages are also compared to simple integrated totals of the (probabilistic) voxel tissue volumes in litres.

[2]We use BW01 to denote the original BrainWeb data `http://www.bic.mni.mcgill.ca/brainweb/selection_normal.html`, and BW04, etc. to denote subject numbers from the 20 new anatomical models `http://www.bic.mni.mcgill.ca/brainweb/anatomic_normal_20.html`.

The effect of varying patient group composition

To provide a clearer characterisation of the impact of atrophy severity on mask construction, the second experiment considers different subsets from a typical VBM cohort of 19 patients with probable Alzheimer's Disease (AD) (M:F 9:10, mean age 68.8) and 19 healthy controls (M:F 8:11, mean age 68.3). T1-weighted volumetric images with a 24 cm field of view, 256x256 matrix, and 124 contiguous 1.5 mm coronal slices were acquired using a spoiled fast GRASS sequence on a 1.5 T Signa scanner (GE Medical Systems, Milwaukee, WI). Acquisition parameters were as follows: repetition time = 15 ms; echo time = 5.4 ms; flip angle = 15°; inversion time = 650 ms. All subjects were recruited from the Cognitive Disorders Clinic at the National Hospital for Neurology and Neurosurgery, London, UK, and gave written informed consent. They were assessed using standard diagnostic criteria. The study was approved by the local ethics committee. For further details see Schott et al. (2005).

This experiment focusses on the robustness of the generated masks with respect to changes in the composition of the subject group, we begin with only the 19 controls, before choosing (based on visual inspection of the scans) a single severely atrophied AD patient to add before re-creating the masks, then finally the 18 remaining AD patients are included, providing a typical balanced two-group comparison.

Practical importance on VBM of Fronto-Temporal Dementia

Finally, the potential for overly restrictive masks to exclude potentially interesting findings in the most atrophied structures is highlighted through presentation of the SPM results for a particular VBM study. A group of 14 Fronto-Temporal Dementia (FTD) patients (M:F

7:7, mean age 63.5) with pronounced and focal temporal lobe atrophy, was compared to a group of 22 approximately matched controls (M:F 10:12, mean age 65.8). All subjects were recruited from a specialist dementia clinic and gave written informed consent. They were assessed using standard diagnostic criteria. The study was approved by the local ethics committee. MR protocols were as described for the AD study.

Results are presented using the masking strategy previously standard within our group, and with an example following the new consensus masking strategy, chosen based on qualitative visual evaluation of the suitability of various masks, prior to performing the statistical analysis.

Results and discussion

Simulated images

Figure 1 illustrates typical results for VBM preprocessing using SPM5's unified segmentation model (Ashburner and Friston, 2005). The estimated segmentation is in close agreement with the simulation's underlying model,[3] but the inter-subject correspondence following spatial normalisation is only approximate. This imperfect overlap necessitates smoothing, but we can observe that even after smoothing there could be poor correspondence if the results were binarised with a relatively high threshold.

[Figure 1 about here.]

[3]In fact, one of the most noticeable differences is that SPM's use of prior tissue probability maps has excluded some unrealistic dural 'GM' present in the simulation.

In figure 2, we explore the results from using the thresholds which maximise the Jaccard similarity coefficients between the binary segmentation from the underlying discrete BrainWeb model and the binarised probabilistic segmentations with and without spatial smoothing using an 8mm full-width at half-maximum (FWHM) Gaussian kernel. It can be seen that while the original segmentation can be binarised successfully at a very high threshold, after smoothing the results are visually not optimal even with the new Jaccard-optimal threshold. The threshold must clearly be lowered to include all desired voxels; fig. 2(f) shows the result from a much lower threshold, which, while often employed in VBM within our group, appears here to be far too generous. This apparent generosity should be contrasted with the findings shown later in figures 5 & 7. These results demonstrate the difficulty of finding a truly optimal threshold for a single subject — even on simulated data with known ground truth. It is for this reason that we allow the threshold to be varied in the consensus masking strategy (in addition to the consensus fraction). Note that our new objective masking strategy is based on the average of the segmentations, so it does not need to estimate optimal thresholds for individual images.

[Figure 2 about here.]

Briefly investigating the use of relative thresholding, table 1 compares the values of SPM's 'global' average to the totals from integrating over voxels, with three different sources of input data for four simulated subjects. It is clear that the total value is insensitive to the choice of these preprocessed source images, unlike the global value. Since the total also has the additional merits of being much simpler to interpret clinically, and of not using an arbitrary threshold (the 1/8 of the original mean), it seems preferable to use these totals as

values for deriving proportional masking thresholds.[4] As a quick check of the suitability of these totals for relative threshold masking, table 2 presents the optimal absolute thresholds for four subjects, and the fractions of the global or total values necessary to achieve these thresholds. There is no apparent problem with using the totals, and limited evidence that they in fact have a more consistent relationship with the optimal threshold than the globals.

[Table 1 about here.]

[Table 2 about here.]

Control and AD group composition

Continuing the comparison of SPM's globals with the integrated tissue totals from the previous experiment, figure 3 shows a strong correlation between these two measurements over all 38 subjects in the group of AD patients and matched controls.

[Figure 3 about here.]

A visual example of the range of atrophy present in this subject-group is given in figure 4. On rough inspection, it might appear from the preprocessed images that the spatial normalisation and smoothing has adequately standardised even the most severely atrophied patient.

[Figure 4 about here.]

[4]While not evaluated here, it would also seem reasonable to prefer these more interpretable and stable values when adjusting for global volume in VBM, either through covariates or scaling-factors.

However, figure 5 presents a range of masks generated from four different strategies, on the three differently composed sub-groups. Table 3 gives the corresponding quantified mask volumes. It is clear from row (a) that the default SPM absolute thresholding strategy is very fragile with respect to the inclusion of atrophied patients. Adding the single severe individual results in a noticeably smaller mask, with particular reductions in the frontal and temporal lobes, and 100ml less total volume. The addition of the remaining 18 AD patients causes a further 120ml reduction in mask volume — corresponding to a loss of approximately 15,000 2mm isotropic voxels. Potentially interesting frontal cortex would not be analysed if such a mask was used. By lowering the consensus from SPM's 100% to 70%, the results become dramatically more robust to the inclusion of the patients. Row (b) of the figure shows only visually minor reductions in the mask; the table reveals that the volume loss through adding the severe case is just a tenth of that with the SPM strategy, though this rises to half when the remaining patients are added.

[Figure 5 about here.]

[Table 3 about here.]

The use of relative thresholding should reduce the sensitivity to disease severity, since more severely atrophied patients will have lower global values and hence lower relative thresholds. However, example results (row c) using SPM's relative threshold masking (based on globals) still show a disturbing loss of cortical GM voxels from the mask with one patient, worsening with the additional patients. The overall loss when adding all patients to just the controls is 180ml (over 12% of the volume of the controls-only mask). Row (d) has the most visually appealing masks, derived from a 70% consensus and a threshold relative to

the integrated total volumes. The loss when adding all patients is now less than 2.5% of the original controls-only mask volume. It is self-evident in this experiment that the mask volume lost when adding a patient group to a control group could correspond to clinically-significant tissue loss in the actual control-patient comparison of interest. That this lost mask-volume can coincide with statistically-significant tissue differences is demonstrated in the next experiment.

Finally within this experiment, one potential problem with over-generous masks is demonstrated. In figure 6 some of the most significant voxels fall in regions where the majority of images do not have substantial chance of being genuine GM tissue. It is difficult to conclude confidently whether or not these are false positives, but the low variance and greater residual roughness present at the illustrated voxel certainly cast some doubt on the strength of the finding. On the other hand, it is also possible that cluster peaks for true-positive findings could be shifted outside the brain due to the effect of smoothing. This is a complex issue, which has received some attention in the literature (Reimold et al., 2006; Acosta-Cabronero et al., 2008). Our recommendation is that the initial mask should not be over-generous, but then if clusters are found which appear to spread beyond the analysis region, the contrast image (numerator from the t-statistic) should subsequently be investigated over a larger mask, visually and/or with software such as that proposed by Reimold et al. (2006) (`http://homepages.uni-tuebingen.de/matthias.reimold/mascoi/`), in order to determine more accurately the extent of the effect.

[Figure 6 about here.]

FTD example

The comparison of FTD patients with healthy controls reveals a pattern of tissue loss with focal left temporal lobe atrophy. Unthresholded SPM t-maps are shown in figure 7 (a) and (b); the two masks used for these analyses are overlaid in (c), where it is immediately obvious that the 100% consensus mask has excluded tissue in the temporal lobes, particularly on the left. The difference in volume of these two masks is over 300ml. Most importantly, (d) shows that some of the statistically-significant voxels (pFWE < 0.05) found when using the more reasonable mask will be ignored in the analysis using the standard 100% consensus mask. This lost significant volume amounts to 8.19ml, or over 1000 2mm isotropic voxels, in exactly the areas that these FTD brains are most atrophied.

[Figure 7 about here.]

Other masking strategies

Software for VBM analysis has recently been released as part of the FMRIB Software Library (Smith et al., 2004), these scripts (`http://www.fmrib.ox.ac.uk/fsl/fslvbm/index.html`) implement a different procedure for their mask creation. FSL-VBM includes voxels in the mask if they meet both the following criteria: the maximum tissue probability over all subjects is at least 0.1; the minimum over the subjects is non-zero. Unlike the SPM strategies so far considered, which are derived from the smoothed (and optionally modulated) normalised segmentations, as used for the statistical analysis, FSL's VBM masking strategy is based on unsmoothed and unmodulated segmentations (even when modulated data are analysed).

[Figure 8 about here.]

Examples of this strategy are illustrated for the AD data-set in figure 8. The most noticeable difference is that the use of unsmoothed segmentations leads to a much rougher mask. For correction of FDR or permutation-based FWE control, this roughness is unlikely to be a problem, but it may be detrimental for RFT-based correction of FWE. Worsley et al. (1996b) reported that expressions for RFT thresholding of statistics appeared to be most accurate for convex search regions, and they suggested that convoluted regions with high surface-area to volume ratios offer no advantage in power over smoother regions with larger volumes.

The lower panels of figure 8 show the results of applying FSL's mask inclusion criteria to the smoothed data which is actually analysed. In this case, smoothing leads to the presence of non-zero voxels as far away from the brain edges as the size of the support of the smoothing kernel used.[5] This effectively leaves only the second criterion in place; that the maximum over the segmentations be over 0.1. Now, we note that this criterion is simply a special case of the consensus masking strategy, where the consensus fraction is the reciprocal of the number of images, i.e. only one image (the maximum one for each voxel) need be above the threshold.

[Figure 9 about here.]

Finally, we consider deriving masks from the average of all subjects' smoothed normalised segmentations. This approach has been reported by Duchesne et al.[6] who binarised their average of 3mm FWHM smoothed unmodulated normalised segmentations at a threshold of

[5] In SPM5, the kernel is non-zero for ± 6 standard deviations.

[6] Unpublished manuscript, available online: `http://www.bic.mni.mcgill.ca/users/duchesne/Proc/NI2004a.pdf`.

0.3. Assuming that there is limited skew in the distribution of voxel intensities over subjects (SPM goes further in assuming normality), the arithmetic mean will approximately equal the median. Since the median by definition has 50% of the data beneath it, thresholding the average should be approximately equivalent to the special case of our proposed masking strategy with a consensus of 50% and the same threshold. In figure 9 we compare these two approaches on the AD data, showing almost identical results. As one would expect from the relatively low consensus fraction, there is very strong robustness to the addition of patients to the control group.

An objective mask strategy

We now use the AD/control data-set to evaluate masks that maximise our proposed optimality criterion (1) based on each sub-group's average image. In figure 10, we compare the 'optimal' thresholds to arbitrarily chosen higher and lower thresholds. The volumes of the optimal masks for the three subject groups are: 1.456, 1.456, and 1.444 litres — exhibiting a loss of below 1% with the addition of the AD patients. This provides a simple fully-automatic and operator-independent technique for creating a mask, which could form a reasonable default option. However, the exact balance of the risks of false positives and false negatives, and of other issues including peak-shifting due to smoothing (Reimold et al., 2006) and unreliability of smoothness estimation outside the brain (discussed below) is a subtle and difficult problem. Therefore it may not be possible to give a definitive masking procedure that would be suitable for all data-sets. If the resulting mask is found to be unacceptable, then manual selection of a different threshold, and/or consensus fraction using

the first technique proposed here, would allow expert quality assurance to be imparted.

[Figure 10 about here.]

Further discussion

VBM studies aiming to localise small lesions or patterns of atrophy in finer scale structures require smaller smoothing kernels, due to the matched filter theorem (Worsley et al., 1996a). The chance of losing interesting voxels from a mask created using absolute or relative thresholding with the standard 100% consensus is likely to be even greater with less smoothing. In a single subject with a severely atrophied small structure, greater amounts of smoothing would permit neighbouring tissue to bring the average value at the atrophied voxels above the threshold. However, it should also be noted that finer scale spatial normalisation (Shen and Davatzikos, 2003; Ashburner, 2007) may counterbalance this effect, as atrophied structures can be better warped to match those of the template/average, with the information about their atrophy being transferred to the deformation field.

In the introduction, we mentioned the need for masking in residual smoothness estimation. This issue is known in the research community[7] but seems not to appear in published literature. Numerical instability in the smoothness estimation (Kiebel et al., 1999) within regions of very low intensity, leads to abnormally high estimated roughness (observable in SPM's resels per voxel RPV image in studies that have used no explicit or threshold masking). Overestimated roughness will not invalidate the RFT results, but will make them over-conservative, possibly to the extent that Bonferroni correction is preferable (SPM's FWE p-values are the more significant of the RFT and Bonferroni versions). A related point

[7]E.g. `http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0701&L=SPM&P=8534`

is that very low intensity voxels are non-negative, which might lead to a departure from normality. Ashburner and Friston (2000) suggested that the use of logistic regression might be preferable for this reason, though it has not proven popular to date.

We observed earlier that thresholds specified relative to some global measure of tissue content have the desirable property of being lower for individuals with more severe atrophy. However, modulated normalised images will also tend to yield lower global values for subjects with smaller total brain volumes or total intracranial volumes (TIV), which is undesirable. This is obvious for integrated segmentation totals, but also true of the standard SPM global measure, where smaller TIV will result in a lower initial mean, letting the mean/8 include more background, and hence reducing the secondary mean. If reliable estimates of TIV can be derived (Whitwell et al., 2001) then it seems likely that a TIV-normalised total segment volume[8] would be the best measure on which to base relative thresholds. In our objective masking strategy, the optimality criterion based on correlation with the average image also lowers the threshold as more atrophied subjects are included in the average; this may be the reason that it exhibited the least sensitivity to the composition of the subject group in the AD/control study. It will again be sensitive to TIV, if using modulation, and hence might be better with TIV-normalised images.

Future work could involve extension of the method of automatic threshold selection, and/or selection of an optimal consensus fraction, perhaps using bootstrap methods or cross-validation. It may also be helpful to base masks upon the voxel-wise statistical results,[9]

[8]Under the assumption that the TIV is approximately proportional to the determinant of the affine component of the normalisation, totals from segments modulated only for nonlinear changes, as described here `http://dbm.neuro.uni-jena.de/vbm/segmentation/modulation/` should give similar results.

[9]SPM follows a related procedure for variance component estimation, only pooling over voxels which show main effects above a certain level of statistical significance (Glaser and Friston, 2004).

directly addressing the problem of false-positives in low-variance regions by excluding these voxels.

Conclusions

With many diseases there is a spectrum of severity of focal atrophy; the most vulnerable regions might also be the most likely to have outlying subjects with particularly severe absence of tissue. The standard masking procedure in the SPM software risks missing findings in the most severely atrophied brain regions. It is important to note that the missed atrophy when using overly restrictive masks might not be readily apparent from consideration of the 'glass-brain' maximum intensity projection commonly presented in VBM results. It seems not to be standard practice for VBM papers to present the analysis region resulting from their choice of masking strategy. We would recommend careful checking of the mask, and would argue in favour of this occurring prior to the statistical analysis itself — a practice which is simplified by using the mask-creation strategy recommended here. We would additionally suggest that the masking procedure be reported clearly enough to be reproducible, as we have previously advocated (Ridgway et al., 2008).

In summary, our suggested protocol is to begin with the objective average-based mask, and to check this before estimating the statistical model; if the mask appears unsuitable we recommend expert visual assessment of masks using other thresholds or the consensus-based strategy. After statistical analysis, if significant findings border the mask edges, we would suggest re-evaluation of the contrast images over a more generous mask to clarify the interpretation of these findings. Software is available to implement our proposed consensus-

based and average-based mask-creation techniques.[10]

Acknowledgments

We are grateful to John Ashburner, for helpful comments on the need for masking in residual smoothness estimation, and to Susie Henley and Jonathan Rohrer for helpful discussion and for experimenting with our new mask creation software. We wish to thank the anonymous reviewers, who made several important suggestions. G.R.R. is funded by an EPSRC CASE Studentship, sponsored by GlaxoSmithKline. J.D.W. is supported by a Wellcome Trust Intermediate Clinical Fellowship. N.C.F. acknowledges support from the UK Medical Research Council. The Dementia Research Centre is an Alzheimer's Research Trust Coordinating Centre. This work was undertaken at UCLH/UCL who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme.

[10] `http://www.cs.ucl.ac.uk/staff/g.ridgway/masking`

References

Acosta-Cabronero, J., Williams, G., Pereira, J., Pengas, G., Nestor, P., 2008. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. Neuroimage 39 (4), 1654–1665.

Ashburner, J., Oct 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113.

Ashburner, J., Friston, K. J., Jun. 2000. Voxel-based morphometry–the methods. Neuroimage 11 (6 Pt 1), 805–821.

Ashburner, J., Friston, K. J., Jul. 2005. Unified segmentation. Neuroimage 26 (3), 839–851.

Aubert-Broche, B., Griffin, M., Pike, G., Evans, A., Collins, D., Nov. 2006. Twenty new digital brain phantoms for creation of validation image data bases. Medical Imaging, IEEE Transactions on 25 (11), 1410–1416.

Belmonte, M., Yurgelun-Todd, D., March 2001. Permutation testing made practical for functional magnetic resonance image analysis. Medical Imaging, IEEE Transactions on 20 (3), 243–248.

Cocosco, C., Kollokian, V., Kwan, R., Evans, A., 1997. Brainweb: Online interface to a 3D MRI simulated brain database. NeuroImage 5 (4), S425.

Crum, W., Camara, O., Hill, D., Nov. 2006. Generalized overlap measures for evaluation and validation in medical image analysis. Medical Imaging, IEEE Transactions on 25 (11), 1451–1461.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., Jan 2008. Bayesian decoding of brain images. Neuroimage 39 (1), 181–205.

Genovese, C. R., Lazar, N. A., Nichols, T., Apr 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15 (4), 870–878.

Glaser, D., Friston, K., 2004. Variance Components, in Human Brain Function 2nd Ed. Academic Press, Ch. 9.

Karas, G. B., Burton, E. J., Rombouts, S. A. R. B., van Schijndel, R. A., O'Brien, J. T., Scheltens, P., McKeith, I. G., Williams, D., Ballard, C., Barkhof, F., Apr 2003. A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18 (4), 895–907.

Kiebel, S. J., Poline, J. B., Friston, K. J., Holmes, A. P., Worsley, K. J., Dec 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. Neuroimage 10 (6), 756–766.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. M., Davatzikos, C., Jan 2004. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage 21 (1), 46–57.

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S., Hodges, J. R., Jan 2000. A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. Ann Neurol 47 (1), 36–45.

Nichols, T. E., Holmes, A. P., 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping 15 (1), 1–25.

Reimold, M., Slifstein, M., Heinz, A., Mueller-Schauenburg, W., Bares, R., Jun 2006. Effect of spatial smoothing on t-maps: arguments for going back from t-maps to masked contrast images. J Cereb Blood Flow Metab 26 (6), 751–759.

Ridgway, G. R., Henley, S. M. D., Rohrer, J. D., Scahill, R. I., Warren, J. D., Fox, N. C., May 2008. Ten simple rules for reporting voxel-based morphometry studies. Neuroimage 40 (4), 1429–1435.

Schott, J. M., Price, S. L., Frost, C., Whitwell, J. L., Rossor, M. N., Fox, N. C., Jul 2005. Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months. Neurology 65 (1), 119–124.

Shen, D., Davatzikos, C., Jan 2003. Very high-resolution morphometry using mass-preserving deformations and HAMMER elastic registration. Neuroimage 18 (1), 28–41.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., Matthews, P. M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208–S219.

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Petersen, R. C., Jack, C. R., Feb 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 39 (3), 1186–1197.

Whitwell, J. L., Crum, W. R., Watt, H. C., Fox, N. C., Sep 2001. Normalization of cerebral

volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging. AJNR Am J Neuroradiol 22 (8), 1483–1489.

Worsley, K., Marrett, S., Neelin, P., Evans, A., 1996a. Searching scale space for activation in PET images. Human Brain Mapping 4 (1), 74–90.

Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., Evans, A., et al., 1996b. A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping 4 (1), 58–73.

List of Tables

1 Global and total values 25
2 Optimal thresholds 26
3 Mask volumes for AD example 27

Table 1: Comparison of SPM's 'Global' averages with integrated totals (in litres) for four BrainWeb subjects, based on native GM segmentations, modulated warped segmentations without smoothing, and with 8mm FWHM smoothing.

	Global			Total		
Subject	Native	Mod. Warped	Smooth M.W.	Native	Mod. Warped	Smooth M.W.
BW01	0.788	0.606	0.402	0.911	0.911	0.911
BW04	0.811	0.650	0.430	0.970	0.970	0.970
BW05	0.782	0.586	0.397	0.907	0.907	0.908
BW06	0.751	0.566	0.387	0.888	0.888	0.888

Table 2: Optimal thresholds, in terms of Jaccard similarity coefficient with BrainWeb model, as absolute values, relative fractions of SPM 'Globals' and of Totals in litres, derived from smoothed segmentations.

Subject	Opt. Abs. Thr.	Opt. Rel. G.	Opt. Rel. T
BW01	0.364	0.904	0.400
BW04	0.379	0.881	0.390
BW05	0.373	0.939	0.411
BW06	0.361	0.934	0.407

Table 3: Mask volumes (in litres) for the masks presented in Figure 5. See figure caption for row descriptions.

Method	Controls	Cs + severe	all subjects
(a)	1.79	1.69	1.57
(b)	1.97	1.96	1.90
(c)	1.47	1.40	1.29
(d)	1.63	1.62	1.59

List of Figures

1 Accurate segmentation; approximate normalisation; need to smooth 29
2 Optimal thresholds, changes with smoothing 30
3 Correlation of 'global' and total volumes 31
4 Example subjects and their segmentations 32
5 Masking results 33
6 AD GLM results 34
7 FTD, masks and regions of significance 35
8 FSL-VBM style masks 36
9 Masks derived from the group mean segmentation 37
10 Optimal thresholding of the group average segmentation 38

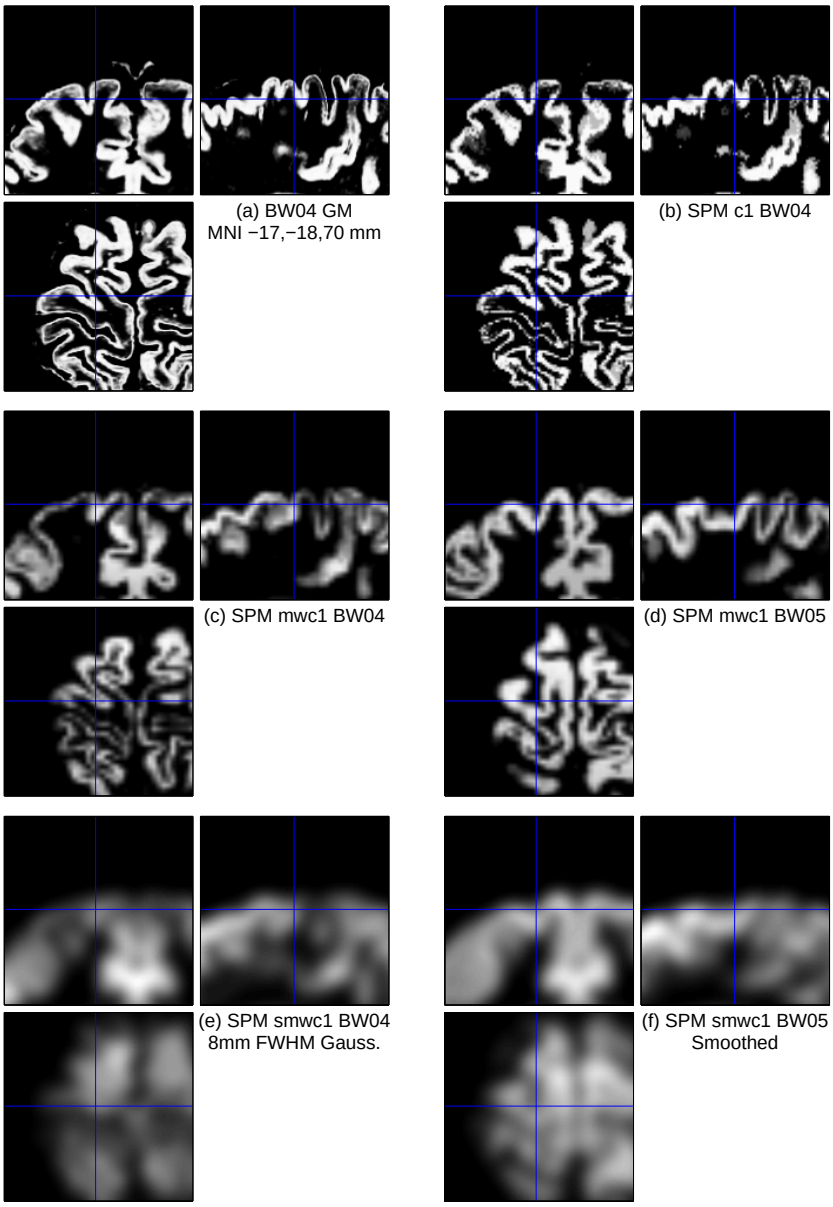

Figure 1: Illustration of the accuracy of tissue segmentation and intersubject spatial normalisation, and the effects of smoothing. (a) and (b) compare the grey matter model used in the BrainWeb simulation to SPM's grey matter segmentation of the simulated T1 image. (c) and (d) show the anatomical correspondence between two different simulated subjects' results after spatial normalisation with a few thousand basis functions. (e) and (f) show the results following spatial smoothing.

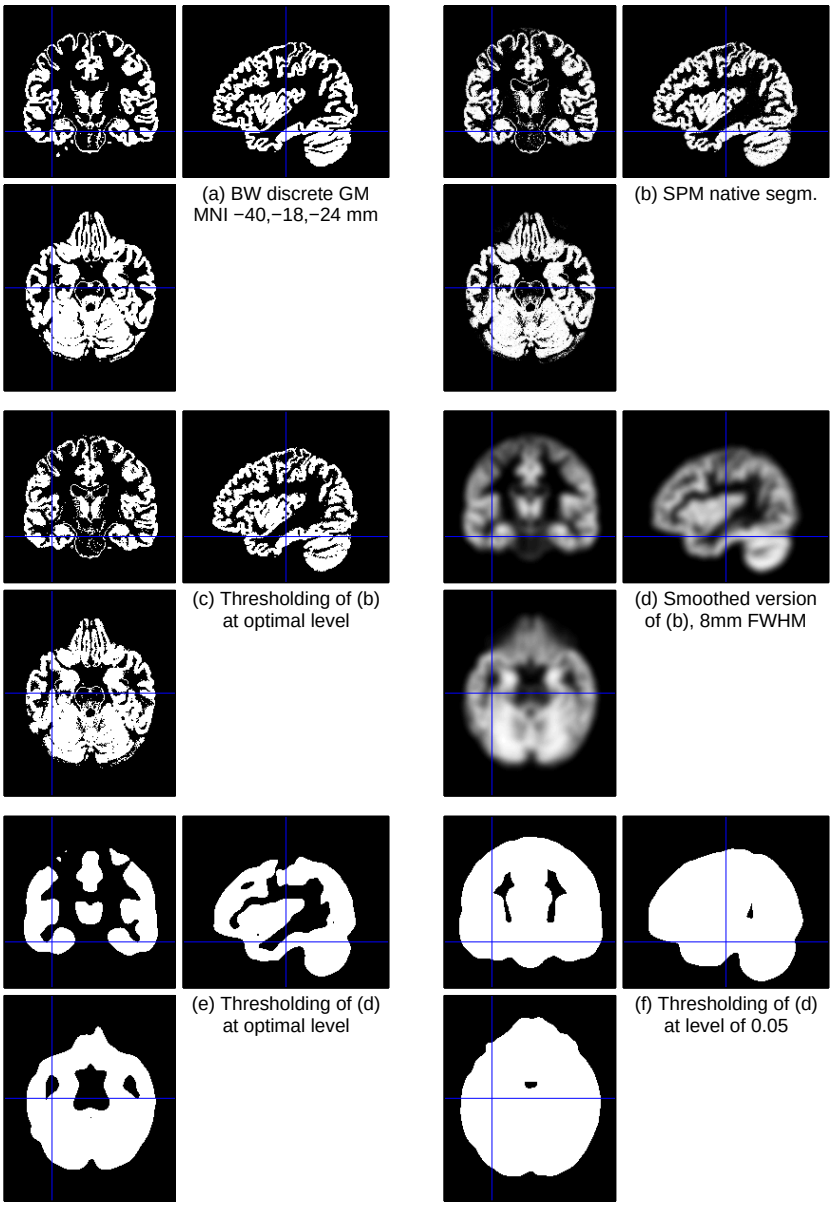

Figure 2: Optimal binarisation of probabilistic segmentations, and the interaction between smoothing and thresholding. (a) The binary GM label for BW01, giving the voxels which have greater probability of being GM than any other tissue. (b) SPM's segmentation of the corresponding simulated T1 image. (c) SPM's segmentation thresholded at a level giving the optimal Jaccard similarity coefficient with the BW01 label. (d) Spatially smoothed SPM segmentation (8mm FWHM Gaussian). (e) and (f) comparison of thresholding of (d) at its 'optimal' level and at a more typical absolute masking threshold.

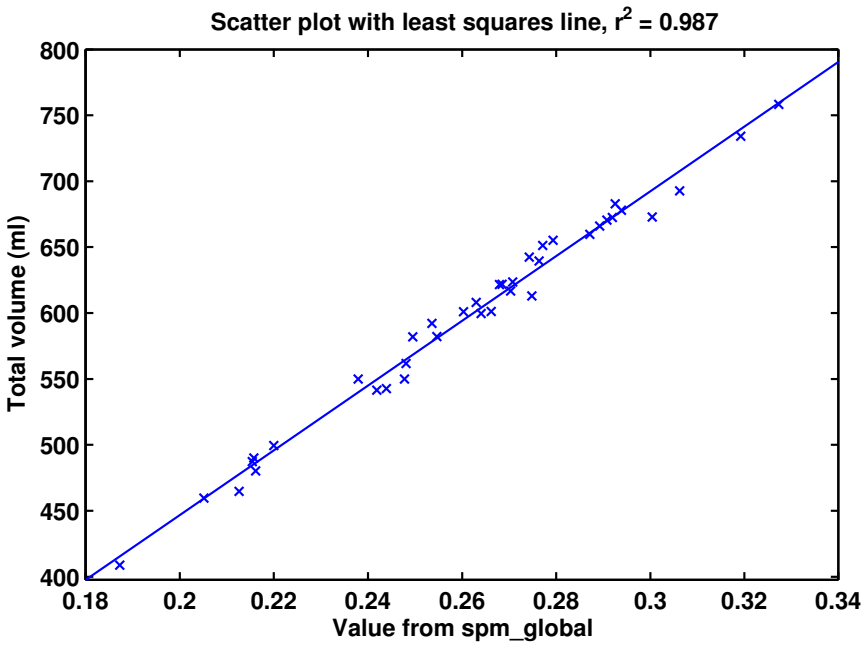

Figure 3: Comparison of total probabilistic GM volumes in ml to SPM's 'global' values. The former come from the summation over all voxels of the segmentation probability multiplied by the voxel volume; the latter come from the mean of the set of voxel segmentation probabilities which exceeded one eighth of the original whole-volume mean.

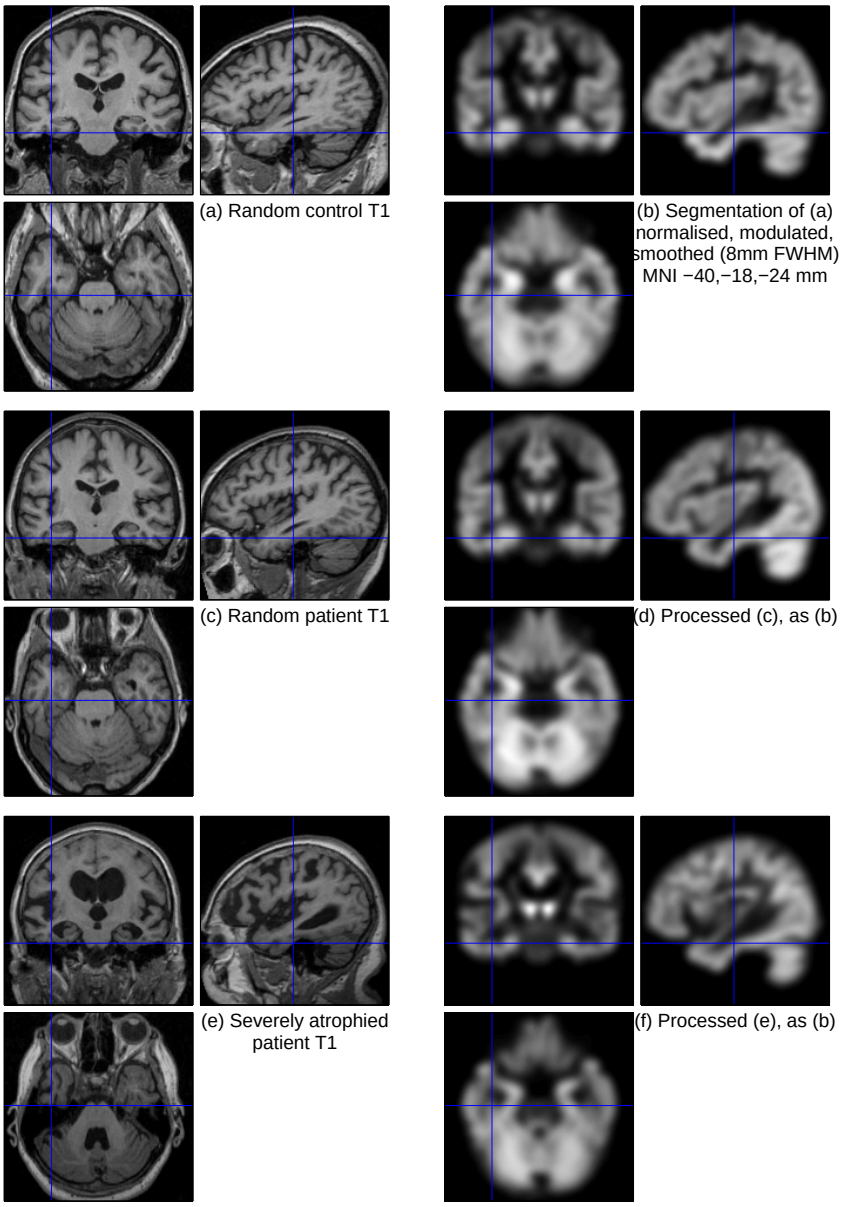

Figure 4: Example subjects. On the left are the standard clinical T1 images; on the right are their corresponding preprocessed segmentations. (a) and (b) are for a typical randomly selected healthy control from the group of 19. (c) and (d) are for one of the 19 AD patients, randomly chosen. (e) and (f) show the most severe AD patient, chosen in terms of visual assessment of overall tissue volume.

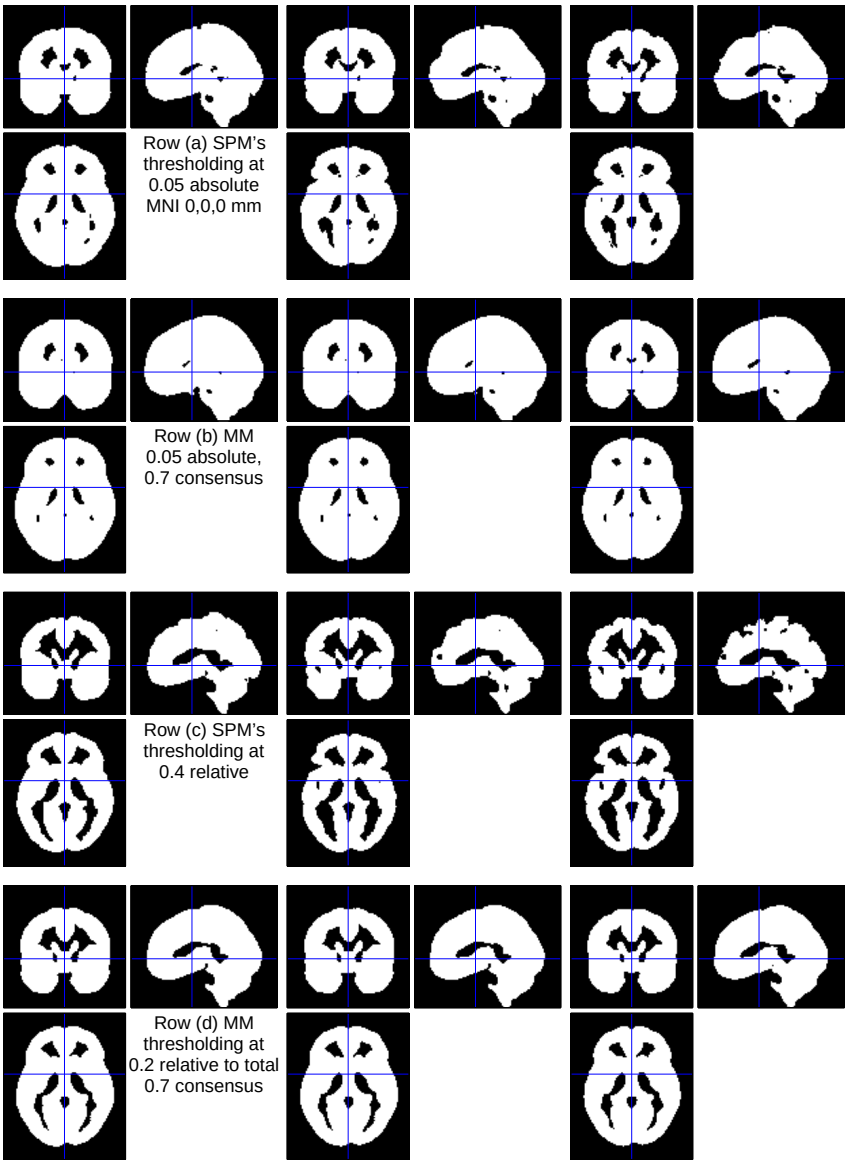

Figure 5: Masking results for varying method and patient group composition. The left column is for the group of 19 healthy controls; middle column, 20 images, including controls and the most severely atrophied patient; right column, entire collection of 19 controls and 19 patients. Rows (a) and (b) present masks based on absolute thresholding at a level of 0.05, in (a) with SPM's default strategy, and in (b) with a "Majority Mask" (MM) requiring 70% of the images over threshold. Rows (c) and (d) investigate relative thresholds. (c) uses SPM's default strategy with thresholds of 0.4 times SPM's global values. (d) requires 70% of the images to exceed thresholds of 0.2 times the total value in litres.

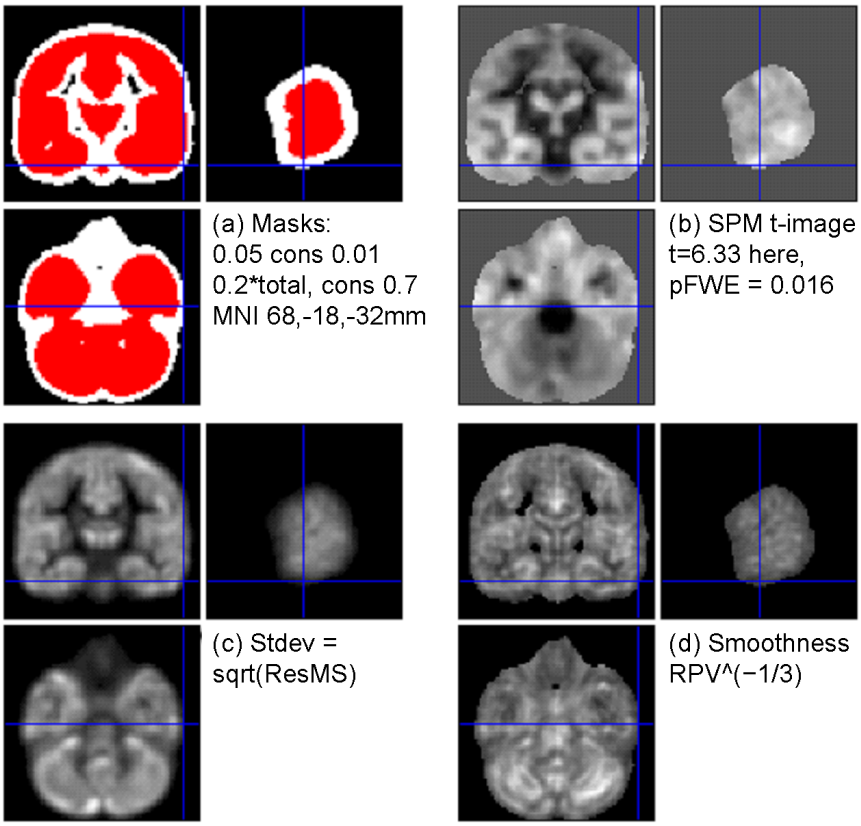

Figure 6: Masks and GLM results for the comparison of controls and AD patients. (a) shows the mask of Fig. 5(d, right column) overlaid on an over-generous mask requiring only one of the images to exceed an absolute threshold of 0.05. (b-d) show the results from GLM estimation using this generous mask, in terms of t-values, standard deviation, and 'smoothness', respectively. The latter is derived from the 'resels per voxel' image, for easier visual interpretation.

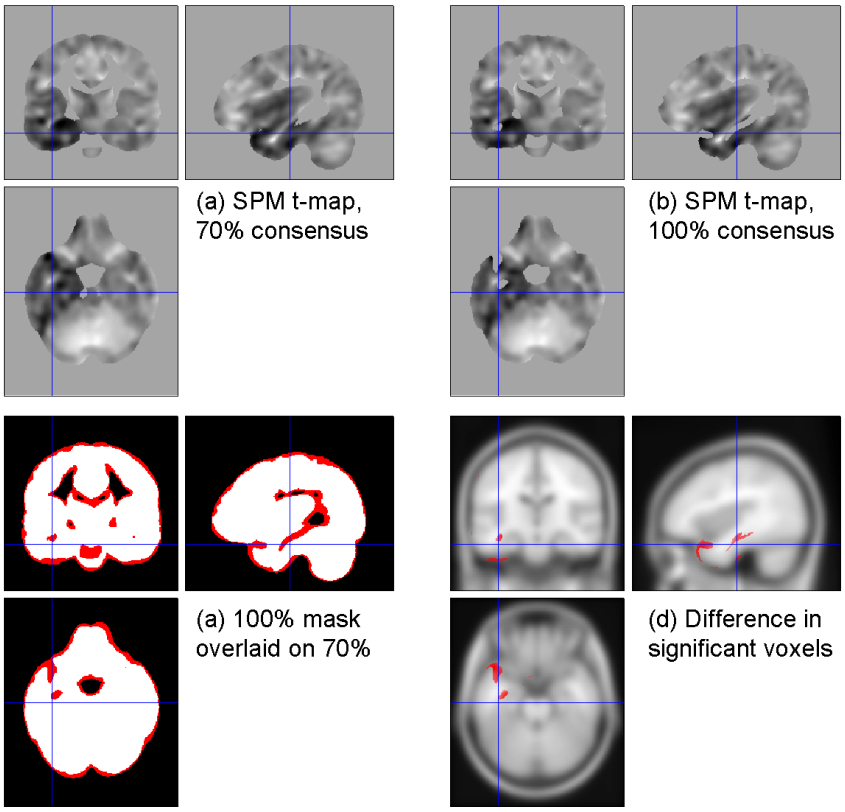

Figure 7: Masks and regions of significance (pFWE < 0.05) for the comparison of FTD subjects with controls. (a) and (b) show t-values for masking requiring either 70% (a) or 100% (b) of images to exceed a threshold of 0.05 (the latter corresponding to SPM's default strategy). (c) overlays the 100% mask on the 70% one. (d) overlaid on the group average segmentation is the region of significance present when using the 70% mask which is excluded from the analysis with the default SPM masking strategy.

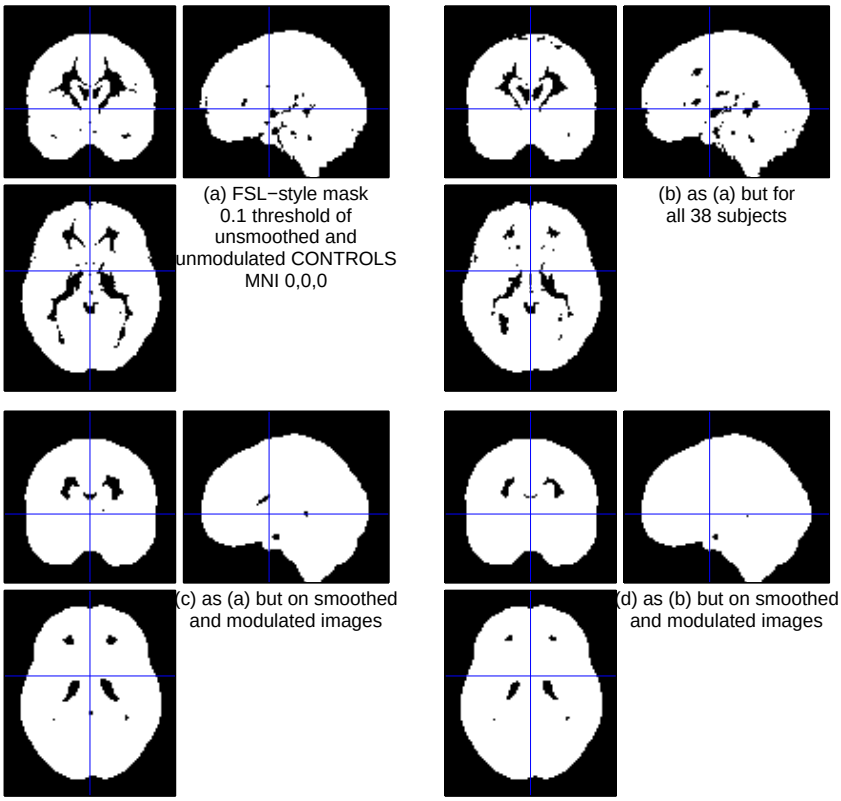

Figure 8: Comparison with the masking strategy used in FSL's VBM implementation. The top row derives masks from unsmoothed and unmodulated normalised segmentations, as in FSL; the bottom row uses smoothed modulated segmentations, as for the other SPM masking strategies discussed here. Left: for controls only; right: for all controls and AD subjects.

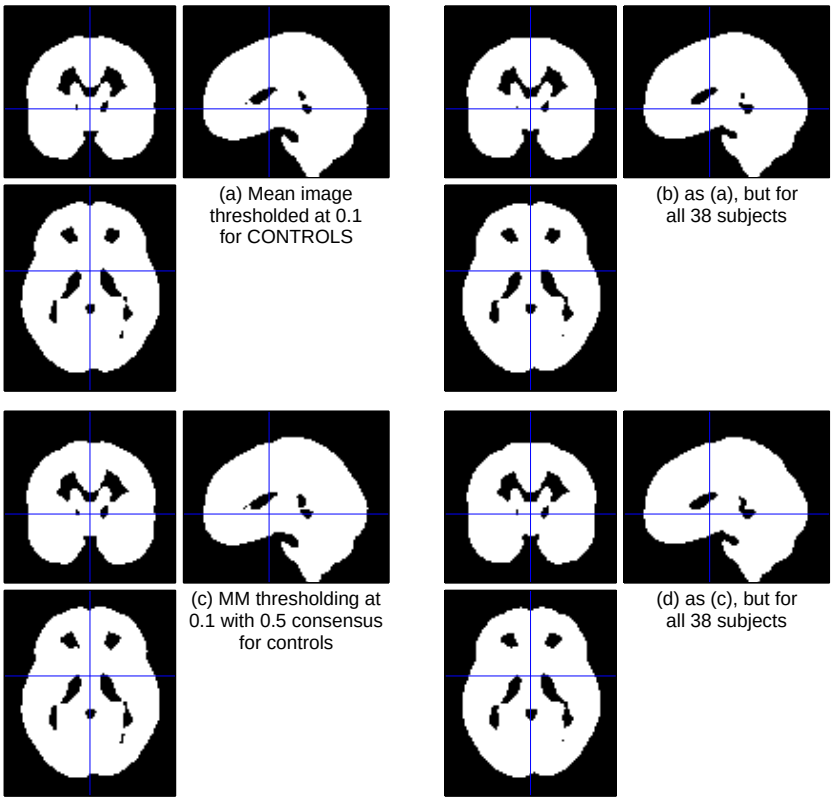

Figure 9: Top row: masks derived from thresholding (at 0.1) the group mean of the smoothed modulated normalised segmentations; bottom row similar masks using a 50% consensus of the unaveraged segmentations and the same threshold. Left: for controls only; right: for all controls and AD subjects.

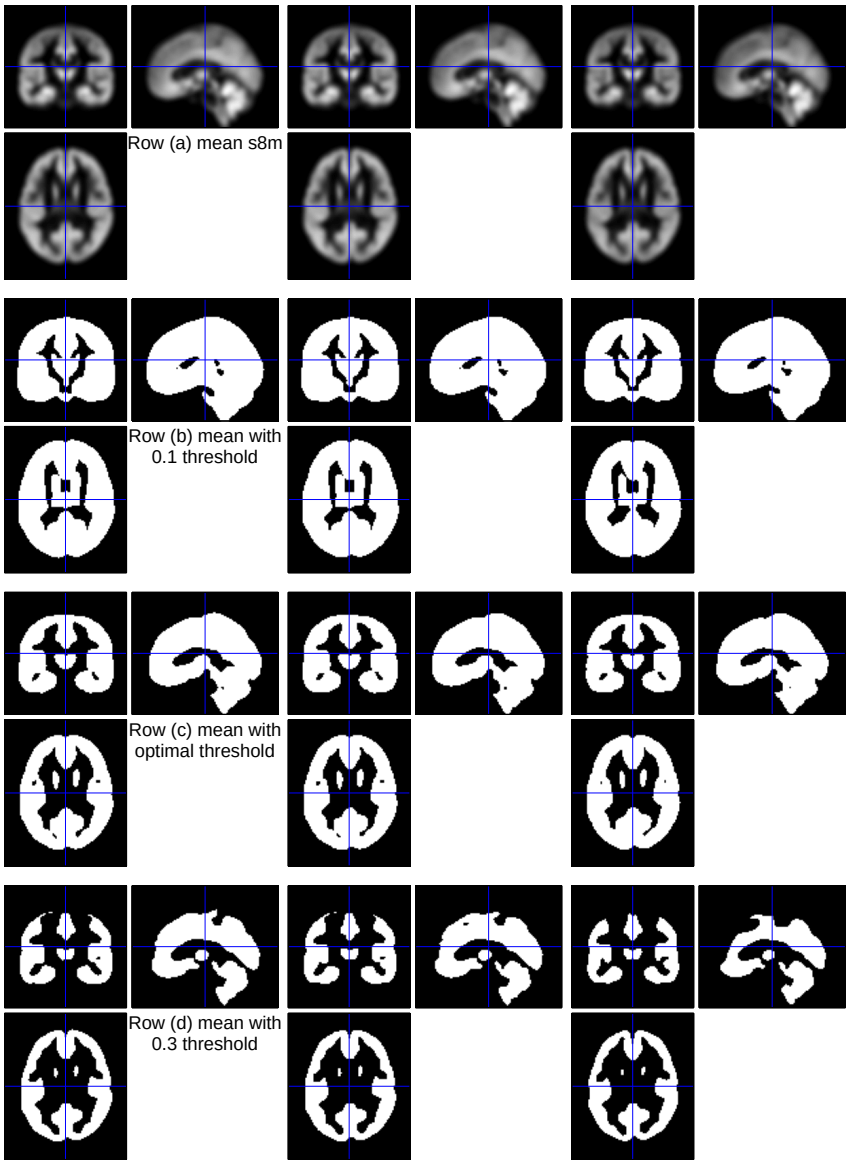

Figure 10: Further investigation of masks derived from the group average segmentation. Left column, for the control group; middle column, controls plus one severe AD patient; right column, all controls and AD subjects. Top row, the average segmentations themselves; row (b) the means thresholded at 0.1 (as in Fig. 9); row (c) the mean images thresholded at optimal levels of 0.203, 0.200, 0.189; bottom row, a higher than optimal threshold of 0.3.