# Longitudinal Voxel-Based Morphometry with Unified Segmentation: Evaluation on Simulated Alzheimer's Disease

Gerard R. Ridgway[a], Oscar Camara[a], Rachael I. Scahill[b], William R. Crum[a],
Brandon Whitcher[c], Nick C. Fox[b], and Derek L. G. Hill[a*]

[a]Centre for Medical Image Computing, University College London.
[b]Dementia Research Centre, UCL. [c]GlaxoSmithKline.

**Abstract.** The goal of this work is to evaluate Voxel-Based Morphometry and three longitudinally-tailored methods of VBM. We use a cohort of simulated images produced by deforming original scans using a Finite Element Method, guided to emulate Alzheimer-like changes. The simulated images provide quite realistic data with a known pattern of spatial atrophy, with which VBM's findings can be meaningfully compared. We believe this is the first evaluation of VBM for which anatomically-plausible 'gold-standard' results are available. The three longitudinal VBM methods have been implemented within the unified segmentation framework of SPM5; one of the techniques is a newly developed procedure, which shows promising potential.

## 1 Introduction

Voxel-Based Morphometry [1] is a method for automated whole-brain analysis of local structural differences, using Statistical Parametric Mapping (`http://www.fil.ion.ucl.ac.uk/spm/`); Longitudinal variants have been developed for application to cohorts with serial imaging [2, 3]. VBM necessitates preprocessing of the images, including spatial normalisation and tissue-segmentation. There is great difficulty in evaluating the performance of VBM methods due to the lack of ground truth. To the best of our knowledge, no previously published VBM studies of realistically complex data have had gold-standard maps of the regions that should be detected.

We have developed Finite Element Methods (FEM) which can structurally alter images, producing finely-controllable, clinically realistic changes [4]. Such simulated images have known underlying deformation fields and volume changes, which can form a gold standard for evaluating atrophy-measurement techniques.

Alzheimer's Disease is a progressive neurodegenerative disorder, of great clinical and socio-economic importance. AD causes a loss of brain tissue which can be visualised and quantified using serial Magnetic Resonance Imaging [5]. Using a cohort of AD patients with MR images at baseline and one year later, we simulated new approximate year-on scans from the original baselines, guided by semi-automated measures of whole-brain, hippocampal, and ventricular volume changes [6]. The original baseline and simulated follow-up images then constitute a data-set with known FEM ground truth; we use this to derive a gold standard suitable for evaluating longitudinal VBM, and compare four such techniques, one of which is novel.

## 2 Methods

### 2.1 Voxel-Based Morphometry and Longitudinal VBM

VBM in SPM5 involves unified tissue-segmentation and spatial normalisation [7], followed by spatial smoothing and voxel-wise statistical testing. With serial data, statistical analysis can take advantage of reduced within-subject variability. To capitalise on the longitudinal information, changes should also be made to the VBM preprocessing methods. In this work, we evaluate standard VBM against two longitudinal methods from the literature (which we have adapted to be compatible with the unified segmentation framework) and our own newly developed SPM5 method.

All SPM analyses were performed within an explicit mask derived from the smoothed ground-truth grey-matter segmentation. All smoothing was done with an 8mm FWHM Gaussian kernel. A one-sample $t$-test was performed on subtraction images; single-tailed contrasts for atrophy (increase$<$0) and 'reverse contrast' of tissue-gain (increase$>$0) were evaluated and thresholded with multiple comparison correction using Random Field Theory ($p_{FWE} < 0.01$).

[*] Derek.Hill@ucl.ac.uk

### 2.1.1 Standard VBM

Here, 'Standard' VBM refers to simple application of unified preprocessing independently to each scan of each subject; only the statistics differ from the non-serial case. 'Standard' should not be contrasted here to 'optimised' VBM [8], which the unified segmentation model aims to supersede [7].

### 2.1.2 Tied-normalisation

The preprocessing step of spatial normalisation should take advantage of the fact that multiple time-points for a single subject can be registered much more accurately than scans of different subjects, and that initial rigid alignment already reveals a great deal about within-subject change [5]. Using the non-unified model of SPM2, Gaser (in Draganski et al. [2]) developed a method with longitudinally tied spatial normalisation, in which repeat scans are transformed using the parameters determined for their corresponding baselines, then independently segmented.

Following the introduction of SPM5's unified framework, an extended generative model for unified longitudinal segmentation and normalisation should ideally be developed. As a simpler alternative, we have implemented an approach which applies the baseline normalisation parameters to the native-space baseline and follow-up grey matter images from separate unified segmentations.

### 2.1.3 Pre-averaged

More advanced techniques can combine inter-subject spatial normalisation with precise intra-subject registration using High-Dimensional Warping (HDW). One such method (designed by Ashburner, and implemented in [3]), creates low-noise averaged images of HDW-registered longitudinal sets, before inter-subject spatial normalisation and segmentation in SPM2. (i.e. averaging is 'pre' segmentation.)

We have adapted this approach to the SPM5 framework, with unified segmentation and inter-subject normalisation following the intra-subject warping and averaging. The intra-subject volume changes from HDW must be taken into account to generate the follow-up data, which can be elegantly done by modulating the native-space segmented average-images with the HDW Jacobian fields before applying the predetermined inter-subject transformations. This avoids the interpolation error due to the transformation of the Jacobians in [3].

### 2.1.4 Post-averaged

We propose a technique similar to pre-averaging, but novel, and well-suited to SPM5's unified segmentation. The new method should be better for subjects with large longitudinal change that might not be fully recovered by HDW, as in this case, the pre-averaged images may be too blurred to segment well. Each time-point is first segmented, and SPM5's bias-corrected version is saved; HDW transformations are then determined on the corrected images and applied to their native-space segmentations. The warped segmentations are then averaged; i.e. averaging is 'post' segmentation of sharp original images. Each average segmentation is modulated with the HDW volume changes to create follow-up equivalents, then each set of original and modulated segmentations is spatially normalised with the baseline parameters.

## 2.2 Finite Element Modelling of Atrophy

The atrophy simulation process is based on that described in [4]. It consists of four main steps: (1) Generation of a reference mesh; (2) Warping to a subject-specific mesh; (3) Deformation of the mesh using a FEM solver; (4) Application of the deformations to the baseline image of each subject, to produce a new simulated follow-up image. The reference mesh was built using the BrainWeb atlas labels of these structures [9] (`http://www.bic.mni.mcgill.ca/brainweb/`). The adaptation of the reference mesh to each subject was achieved with a mesh warping procedure guided by a fluid registration algorithm [10].

We used a cohort of 18 probable AD patients (7 female; ages from 55 to 86, mean 70) with baseline and 12-month follow-up MRI scans [6]. The FEM simulation was driven using values of the subjects' volume changes in the brain, hippocampi, and ventricles (from semi-automated segmentation-based measurements). Simulated mean (standard deviation) percentage volume increases were: brain, -2.43 (1.18); hippocampi, -4.74 (3.24); ventricles, 11.49 (5.35). Figure 1(a-c) shows a single-subject example of atrophy simulation; ventricular expansion, cortical thinning, and opening of CSF spaces can be observed.

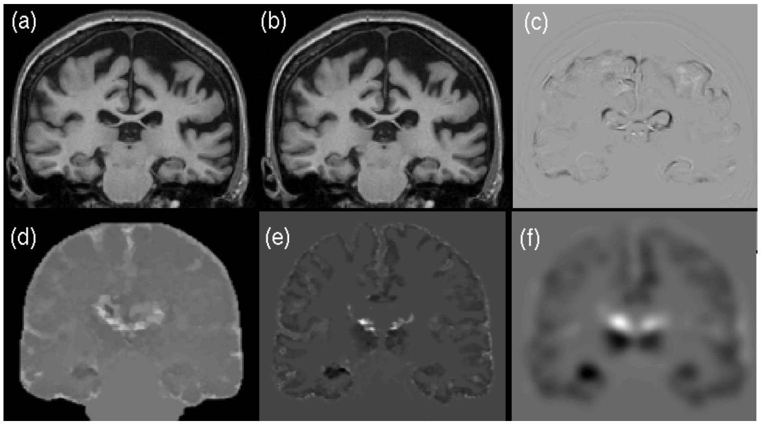

**Figure 1.** Example case of simulated atrophy: (a) Original baseline; (b) Simulated +1yr follow-up; (c) Subtraction image. The same subject's gold-standard volume changes in BrainWeb space: (d) Volume gain (VG=1yr/orig); (e) GM-increase = (GM*VG)-GM; and (f) Smoothed GM-increase, as entered into the analysis.

## 2.3 Generation of a Gold Standard

Because the same mesh is warped from the BrainWeb template to each individual patient, there is a known correspondence between elements of the warped meshes for the different subjects; therefore the volume change of each element can be mapped back to the common space. By converting the element-wise volume changes to a voxel-wise representation, an image of the ratio of follow-up to original volume is created. These volume gain ratio images can be used to modulate the BrainWeb Grey Matter Segmentation, resulting in perfectly aligned effective follow-up segmentations, similar to those in the two HDW-based longitudinal VBM methods. The original BrainWeb GM is then subtracted from each follow-up and the result smoothed. Figure 1(d-f) illustrates this process for one subject. The gold-standard smoothed subtraction images could be entered into an identical one-sample $t$-test as the actual sets of VBM subtraction images. For reasons discussed in section 4.1, we instead use contrast (negative mean) images, thresholded at different values for visualisation purposes.

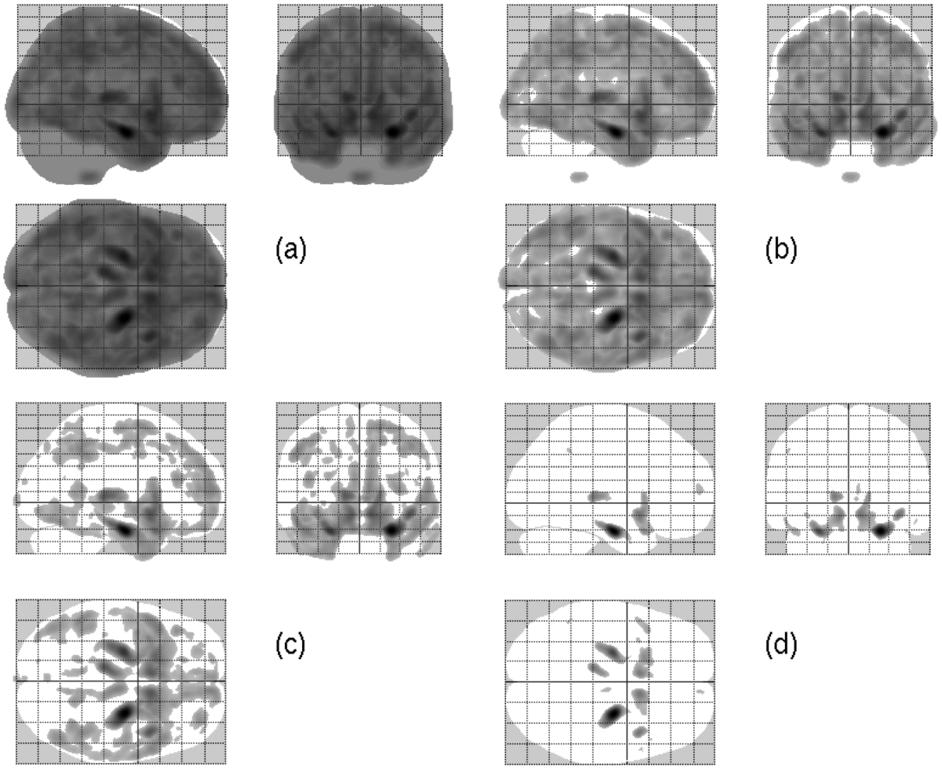

**Figure 2.** Gold-standard average atrophy, Maximum Intensity Projections thresholded at: (a) 0, (b) 0.01, (c) 0.02, (d) 0.03.

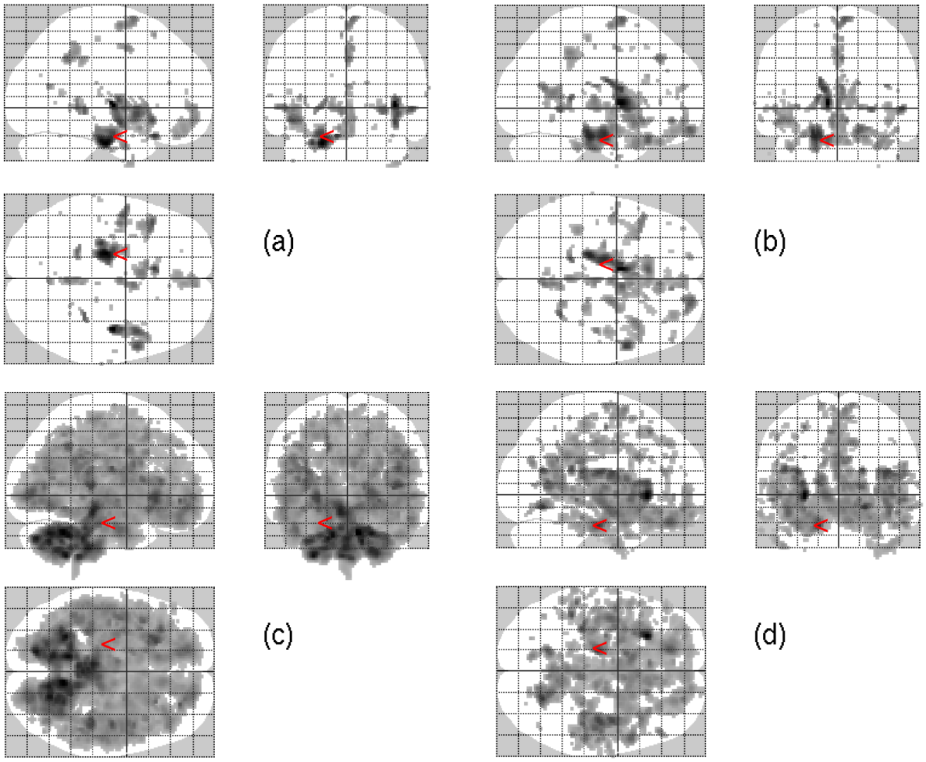

**Figure 3.** Maximum Intensity Projections of significant atrophy ($p_{FWE} < 0.01$) for VBM methods: (a) Standard; (b) Tied-normalisation; (c) Pre-averaged; (d) Post-averaged.

## 3 Results

Gold-standard maximum intensity projections, at varying thresholds, can be seen in figure 2. Statistical results from the four VBM methods are presented in figure 3 as maximum intensity projections. In both cases the atrophy $t$-contrast (increase$<$0) is shown. For the 'reverse contrast' (i.e. gain of GM over time), none of the four methods detected any voxels at the corrected level. Table 1 shows correlations between the ground truth contrast image and the contrast or $t$-value images for the four methods:

| Method | std | tied | pre | post |
|---|---|---|---|---|
| contrast | 0.24 | 0.17 | 0.63 | 0.65 |
| $t$-values | 0.16 | 0.10 | 0.47 | 0.36 |

**Table 1.** Image-wise Spearman rank-correlations (over in-mask voxels).

## 4 Discussion

### 4.1 From Simulation Ground Truth to VBM Gold Standard

The atrophy simulation method gives the ground-truth volume change over the nodes of the deformed mesh. However, several steps are required to convert this data into a gold standard for VBM. Following the obvious approach of performing the same statistical analysis of the ground-truth GM-increase images as of the VBM subtraction images, we obtained unrealistic $t$-maps (not shown). Unreasonably large $t$-values occur outside the regions in which FEM volume changes were introduced. This may be due to the low spatial variance of the volume change maps outside these areas. Inter-subject spatial variability is far lower for these images than it is for natural anatomical variation in real patient images, even after spatial normalisation.

Instead, we threshold the 'contrast image' — the numerator of the $t$-statistic. The maximum intensity projections (see figure 2) now look sensible, but this method leaves open the question of at what level the contrast image should be thresholded. On the other hand, there is also a degree of arbitrariness in the choice of threshold ($\alpha$) for statistical tests. The approach taken here of presenting differently-thresholded versions of the gold standard allows a multi-scale evaluation of the pattern of atrophy. Note that the purpose of the gold standard is to indicate the spatial pattern of simulated atrophy; its significance is not of innate interest.

### 4.2 Longitudinal VBM

The gold-standard results shown in figure 2 indicate the presence of diffuse global atrophy, with greater focus on the temporal lobes and strongest change in the hippocampi; the cerebellum and brain-stem are spared. The key advance of this work is that the VBM methods may be compared directly to this desired pattern, as well as to each other.

We first note that all four methods appear to perform better at detecting the hippocampal and temporal lobe atrophy compared to the more diffuse cortical atrophy. This is probably due to the greater natural anatomical variation in the pattern of cortical folding. Standard VBM detects the least atrophy of the four methods, though there are no obvious false-positive regions. Longitudinally tied normalisation seems to give only minor improvements, though there is some evidence that the less variable tied registration preserves more of the cortical atrophy. Intriguingly, the correlation with ground truth (table 1) appears worse. Both HDW-based methods appear much more sensitive, though some of the atrophy they report is not apparently well-matched to the gold standard (e.g. the insula). In addition, some areas present in the gold standard appear to be missed despite the greater apparent sensitivity (e.g. temporal horns, and the focal nature of the hippocampal atrophy). The correlations in table 1 reaffirm the superiority of the HDW-based methods, but don't allow a clear preference for either.

The pre-averaging method [3] seems to produce false-positive results in the cerebellum and brainstem. Our new post-averaging method appears to avoid this, at the expense of detecting less true cortical atrophy. Additionally, the post-averaging method has better detected the hippocampal atrophy. Reasons for the differences are not entirely clear, as both methods use the same HDW transformations. The pre-averaging method segments (and normalises) an image with higher signal-to-noise ratio but potentially significant blurring; while post-averaging of original (lower SNR segmentations) also improves the SNR of the results. The relative merits of the two alternatives need further investigation.

We note that there are statistical objections to the comparison of $t$- or $p$-values, as a difference in significance is not equivalent to a significant difference. However, with fundamentally different methods such as the standard and HDW-based VBM evaluated here, there is a risk of registration problems if the ANOVA interactions between atrophy and method are tested. The findings shown here are intended to allow comparison between distributions of detected atrophy of the methods and gold standard, with the aims of informing choices between different VBM methods, and guiding further comparative studies. In the future, we hope to perform more quantitative analysis, including a direct ground-truth based investigation of segmentation performance. The evaluation may be complemented by testing the methods on real data.

## Acknowledgements

GRR is supported by an EPSRC CASE Studentship sponsored by GSK. RIS and NCF are supported by the MRC.

## References

1. J. Ashburner & K. J. Friston. "Voxel-based morphometry–the methods." *Neuroimage* **11(6 Pt 1)**, pp. 805–821, June 2000.
2. B. Draganski, C. Gaser, V. Busch et al. "Neuroplasticity: changes in grey matter induced by training." *Nature* **427(6972)**, pp. 311–312, Jan 2004.
3. G. Chételat, B. Landeau, F. Eustache et al. "Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study." *Neuroimage* **27(4)**, pp. 934–946, Oct 2005.
4. O. Camara, M. Schweiger, R. Scahill et al. "Phenomenological Model of Diffuse Global and Regional Atrophy Using Finite-Element Methods." *Medical Imaging, IEEE Transactions on* **25(11)**, pp. 1417–1430, Nov. 2006.
5. N. C. Fox & J. M. Schott. "Imaging cerebral atrophy: normal ageing to Alzheimer's disease." *Lancet* **363(9406)**, pp. 392–394, Jan 2004.
6. J. M. Schott, S. L. Price, C. Frost et al. "Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months." *Neurology* **65(1)**, pp. 119–124, Jul 2005.
7. J. Ashburner & K. J. Friston. "Unified segmentation." *Neuroimage* **26(3)**, pp. 839–851, July 2005.
8. C. D. Good, I. S. Johnsrude, J. Ashburner et al. "A voxel-based morphometric study of ageing in 465 normal adult human brains." *Neuroimage* **14(1 Pt 1)**, pp. 21–36, July 2001.
9. D. Collins, A. Zijdenbos, V. Kollokian et al. "Design and construction of a realistic digital brain phantom." *Medical Imaging, IEEE Transactions on* **17(3)**, pp. 463–468, June 1998.
10. W. R. Crum, C. Tanner & D. J. Hawkes. "Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging." *Phys Med Biol* **50(21)**, pp. 5153–5174, Nov 2005.