Toolbox

Editor's Note: Toolboxes are intended to briefly highlight a new method or a resource of general use in neuroscience or to critically analyze existing approaches or methods. For more information, see http://www.jneurosci.org/misc/itoa.shtml.

Voxel-Based Morphometry: An Automated Technique for Assessing Structural Changes in the Brain

Jennifer L. Whitwell
Department of Radiology, Mayo Clinic, Rochester, Minnesota 55905

Introduction

The identification of structural changes in the brain on magnetic resonance imaging (MRI) scans is increasingly important in the study of neurological and psychiatric diseases. MRI can be used to identify and exclude treatable causes of cognitive impairment and it has also become important in the differential diagnosis of disease, in tracking disease progression, and for research purposes. Pathological changes in the brain resulting in cell loss manifest as loss of brain tissue, or atrophy, which can be detected by structural MRI. Characteristic patterns of atrophy are associated with specific neurodegenerative diseases. Traditional techniques of analyzing atrophy on MRI include visual assessment by experienced radiologists and manual measurements of structures of interest. However, automated techniques have been developed which allow the assessment of atrophy across large groups of subjects without the need for time-consuming manual measurements or subjective visual assessments.

Voxel-based morphometry (VBM) is one such automated technique that has grown in popularity since its introduction (Wright et al., 1995; Ashburner and Friston, 2000), largely because of the fact that it is relatively easy to use and has provided biologically plausible results. It uses statistics to identify differences in brain anatomy between groups of subjects, which in turn can be used to infer the presence of atrophy or, less commonly, tissue expansion in subjects with disease. The technique typically uses T1-weighted volumetric MRI scans and essentially performs statistical tests across all voxels in the image to identify volume differences between groups. For example, to identify differences in patterns of regional anatomy between groups of subjects, a series of *t* tests can be performed at every voxel in the image. Regression analyses can also be performed across voxels to assess neuroanatomical correlates of cognitive or behavioral deficits. The technique has been applied to a number of different disorders, including neurodegenerative diseases (Whitwell and Jack, 2005), movement disorders (Whitwell and Josephs, 2007), epilepsy (Keller and Roberts, 2008), multiple sclerosis (Prinster et al., 2006; Sepulcre et al., 2006), and schizophrenia (Williams, 2008), contributing to the understanding of how the brain changes in these disorders and how brain changes relate to characteristic clinical features. Although results from VBM studies are generally difficult to validate, studies have compared results of VBM analyses to manual and visual measurements of particular structures and have shown relatively good correspondence between the techniques (Good et al., 2002; Giuliani et al., 2005; Whitwell et al., 2005; Davies et al., 2009), providing some confidence in the biological validity of VBM.

MRI processing

In order for statistical analyses to be performed across multiple MRI scans from different individuals, the MRI scans need to be matched together spatially (i.e., registered) so that a location in one subject's MRI corresponds to the same location in another subject's MRI. This process is known as spatial normalization. This is not an easy thing to do given that anatomy varies a great deal across subjects, and heads will be in different positions in the scanner. It is generally achieved by registering all images from a study onto the same template image so that they are all in the same space. Different algorithms can be used to perform this registration (Ashburner and Friston, 2000; Davatzikos et al., 2001), but they typically include a nonlinear transformation (Ashburner and Friston, 2000). The most commonly applied algorithm available in the Statistical Parametric Mapping (SPM) software involves performing a 12-parameter affine transformation followed by a nonlinear registration using a mean squared difference matching function (Ashburner and Friston, 2000). The template image used for the spatial normalization could be one specific MRI scan or could be created by averaging across a number of different MRI scans that have been put in the same space. Customized templates that are created using the study cohort or a cohort that is matched to the study cohort in terms of age, disease status, scanner field strength, and scanning parameters are recommended for registrations that use a mean squared difference matching function to improve the normalization between each subject in the study cohort and the template (Good et al., 2001; Senjem et al., 2005).

Received May 7, 2009; revised June 22, 2009; accepted June 26, 2009.

I acknowledge Dr. Keith A. Josephs and Dr. Clifford R. Jack for their helpful comments and suggestions.

Correspondence should be addressed to Dr. Jennifer L. Whitwell, Assistant Professor of Radiology, Department of Radiology, Mayo Clinic, 200 1st St SW, Rochester, MN 55905. E-mail: whitwell.jennifer@mayo.edu.

DOI:10.1523/JNEUROSCI.2160-09.2009

 0270-6474/09/299661-04$15.00/0

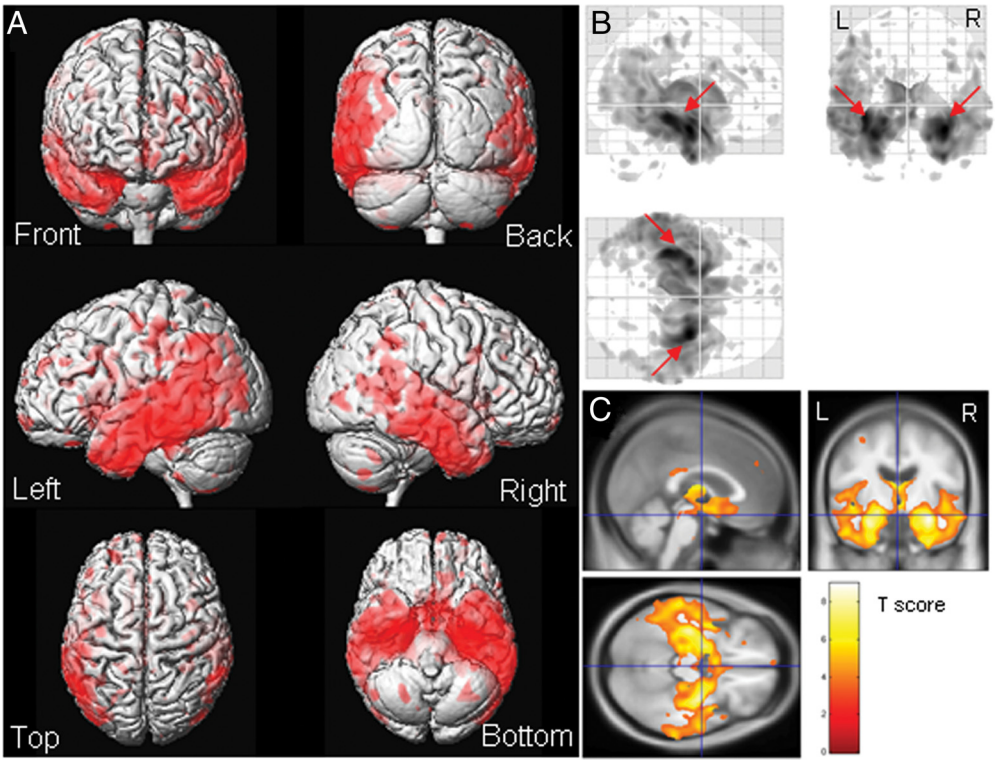

Figure 1. Different VBM display options. Results of a VBM *t* test analysis assessing patterns of gray matter loss in a group of 20 subjects with a clinical diagnosis of Alzheimer's disease (AD) compared with a group of 40 healthy control subjects. The healthy control subjects have been matched by age and gender to the subjects with Alzheimer's disease. The analysis was performed with SPM2 using a statistical threshold of $p < 0.005$ after correction for multiple comparisons using the FDR correction method. The same results have been shown using three different display techniques. ***A***, The voxels that showed significantly reduced gray matter volume in the AD subjects compared with controls are shown in red on a 3D render of a brain. Six different views of the render are shown. ***B***, The voxels that showed significantly reduced gray matter volume in the AD subjects compared with controls are shown as grayscale on a glass-brain render of the brain. Sagittal, coronal, and axial views are shown. ***C***, The voxels that showed significantly reduced gray matter volume in the AD subjects compared with controls are shown on representative sagittal, coronal, and axial slices through the template image with the color bar representing the *t* statistic. The 3D renders (***A***) demonstrate that gray matter loss was present predominantly in the temporoparietal cortex in the AD subjects. The glass-brain renders (***B***) provide the same information but also provide information on the regions of greatest loss, showing that the regions of most severe loss are located in the temporal lobes (arrows). The slices through the template (***C***) provide more detail on exactly which structures of the brain show gray matter loss, yet only show regions of loss on the selected slices. They demonstrate that gray matter loss was observed bilaterally throughout the temporal lobes, particularly involving the hippocampus, and also in the insula and posterior cingulate. All three display techniques therefore provide complementary information.

Images are segmented into different tissue compartments (gray matter, white matter, and CSF), and analysis is performed separately on either gray or white matter, dependent on the question being asked. There are a number of ways to perform the segmentation, including using prior probability maps as well as voxel intensity to guide segmentation, as in SPM. Such prior probability maps may be more unbiased when generated from the specific population under study. In SPM, in which a low-parameter shape transformation is performed for spatial normalization, a step called modulation is then often applied which aims to correct for volume change during the spatial normalization step (Good et al., 2001). Image intensities are scaled by the amount of contraction that has occurred during spatial normalization, so that the total amount of gray matter remains the same as in the original image. The analysis will then compare volumetric differences between scans. If the spatial normalization was precise, and all the segmented images appeared identical, no significant differences would be detected in unmodulated data, and the analysis would reflect registration error rather than volume differences. Other techniques that use different normalization procedures, such as the RAVENS (regional volumetric analysis of brain images) method, which uses high dimensional elastic transformations using point correspondence (Davatzikos, 1998; Davatzikos et al., 2001), preserve the volume of different tissues and so do not require a separate modulation step.

Finally, the images are smoothed (Ashburner and Friston, 2000; Good et al., 2001) whereby the intensity of each voxel is replaced by the weighted average of the surrounding voxels, in essence blurring the segmented image. The number of voxels averaged at each point is determined by the size of the smoothing kernel, which can vary across studies (Rosen et al., 2002; Karas et al., 2003; Whitwell et al., 2009). Smoothing makes the data conform more closely to the Gaussian field model, which is an important assumption of VBM, renders the data more normally distributed, increasing the validity of parametric tests, and reduces intersubject variability (Ashburner and Friston, 2000; Salmond et al., 2002). Smoothing increases the sensitivity to detect changes by reducing the variance across subjects, although excessive smoothing will diminish the ability to accurately localize change in the brain.

Although these processing steps are necessary to be able to analyze data across subjects, they can also introduce errors and variability into the analysis, which can reduce sensitivity. For example, VBM cannot differentiate real changes in tissue volume from local mis-registration of images (Ashburner and Friston, 2001; Bookstein, 2001). Normalization accuracy will vary across regions and, therefore, the ability to detect change will differ across regions. The accuracy of the segmentation will also depend on the quality of the normalization. Iterative normalization and

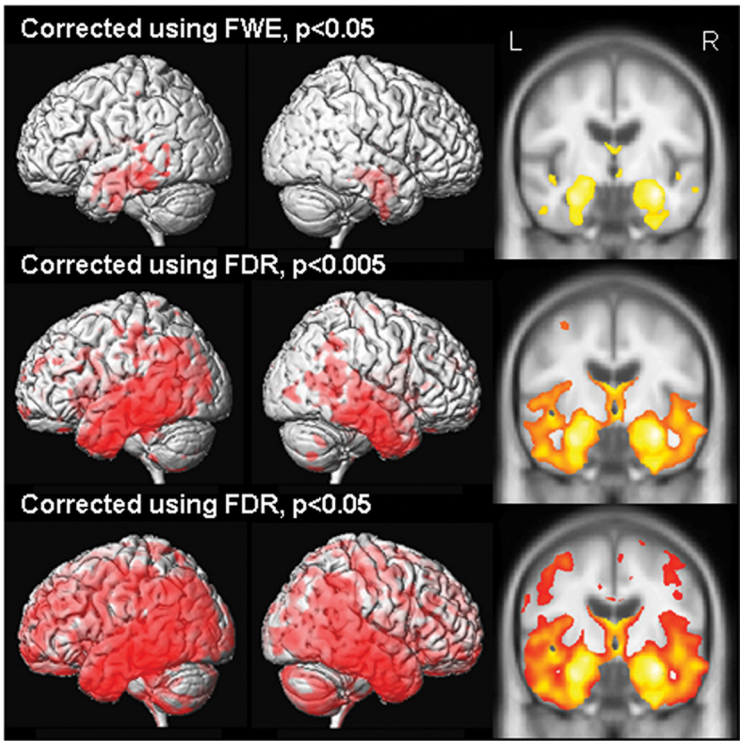

Figure 2. Effect of statistical threshold choice. Results of a VBM *t* test analysis assessing patterns of gray matter loss in a group of 20 subjects with a clinical diagnosis of Alzheimer's disease (AD) compared with a group of 40 healthy control subjects, shown at different statistical thresholds. The top analysis uses a statistical threshold of $p < 0.05$ after correction for multiple comparisons using the FWE correction. This analysis has a very low probability of false-positive results, yet has a greater chance of false negatives and failing to identify regions of the brain that may truly be atrophic. Gray matter loss in this analysis was identified exclusively in the temporal lobes, particularly in the medial temporal lobes. The middle analysis shows the results after correction for multiple comparisons using the more lenient FDR correction at $p < 0.005$ and shows a more widespread pattern of gray matter loss involving the temporal lobes, but also the parietal and frontal lobes. The bottom analysis shows the most statistically lenient analysis after correction using the FDR at $p < 0.05$. In this analysis gray matter loss is observed throughout the entire brain, including regions that are typically spared in AD such as the cerebellum and sensorimotor cortices, which likely represent false positives. Therefore, the statistical threshold chosen for the analysis will influence the resultant pattern of gray matter loss and, potentially, the conclusions of the studies concerning which regions are affected in the disease.

segmentation methods have been developed which aim to optimize both procedures concurrently to improve the final segmentations (Ashburner and Friston, 2005). Segmentation errors can also occur because of displacement of tissue and partial volume effects between gray matter and CSF, which are both especially likely to occur in atrophic brains. The use of customized templates can help to minimize some of these potential errors (Good et al., 2001).

Statistical analysis of VBM results

Statistical analysis of the smoothed segmented images can be performed with parametric statistics using the general linear model and the theory of Gaussian random fields to ascertain significance (Ashburner and Friston, 2000), although nonparametric testing can also been applied (Nichols and Holmes, 2002; Ziolko et al., 2006; Rorden et al., 2007). The null hypothesis is that there is no difference in tissue volume between the groups in question. These analyses generate statistical maps showing all voxels of the brain that refute the null and show significance to a certain, user-selected, *p* value. These maps are often shown as color maps with the scale representing the *t* statistic, but can also be shown as three-dimensional (3D) surface renders of the brain or on what is known as the "glass-brain" display in which all significant voxels are displayed on an essentially transparent render (Fig. 1). Although both gray and white matter volumes can be assessed using VBM, the majority of VBM studies concentrate on gray matter. Changes in white matter integrity may be assessed more accurately using imaging techniques such as diffusion tensor imaging.

Because the statistical tests are performed across a very large number of voxels, it is important that studies correct for multiple comparisons to prevent the occurrence of false positives. There are a couple of typical methods used to perform such a correction, such as the family-wise error (FWE) correction (Friston et al., 1993) and the more lenient false discovery rate (FDR) correction (Genovese et al., 2002), which both reduce the chance of false-positive results (www.fil.ion.ucl.ac.uk). The FWE correction controls the chance of any false positives (as in Bonferroni methods) across the entire volume, whereas the FDR correction controls the expected proportion of false positives among suprathreshold voxels. A number of studies have also used what is called a small volume correction to reduce the number of comparisons being performed and increase the chance of identifying significant results in particular regions of interest. This method typically involves placing regions of interest over particular structures and only performing analysis over these regions. The placement of these regions should be hypothesis driven and ideally based on previous work.

Interpreting VBM results

Interpreting data across VBM studies is a problem because there are a large number of factors that can vary and influence the results. First, the processing steps often vary across studies (Whitwell and Jack, 2005), with studies using different degrees of smoothing and different registration and segmentation algorithms. Second, as well as having different options for correction for multiple comparisons, there are no standard conventions for what *p* value to apply to each statistical analysis, leading to variability across studies. It is important to understand that by changing the *p* value and using different corrections for multiple comparisons, the number of voxels that exceed the significance threshold will change, and this could potentially change the final conclusions of the study (Fig. 2). Studies also vary greatly in the number of subjects included in both control and disease cohorts, which in turn can have a large effect on the resulting *p* values. As with traditional statistical tests, the power to detect differences between groups will typically be a function of the sample size, the degree of the investigated "effect," and the error probability. Therefore, the larger the sample size, the greater the power to detect differences, although differences can be observed with smaller cohorts if the effect size is large. Consequently, studies with larger sample sizes will typically be able to apply the harsh FWE correction for multiple comparisons, whereas smaller studies may favor

the more lenient FDR correction. The resulting power will also depend on errors introduced in the image processing steps and variability across subjects. There are also many potential confounders that can influence the results of a VBM study, for example, differences in age, gender ratios, or disease severity across groups. These potential confounders need to be properly addressed in any study design to be able to make appropriate conclusions concerning the results.

Given all this potential variability, a comparison of *t* statistics or *p* values across studies does not tell us anything biologically meaningful, and only provides anecdotal evidence for differences between diseases and different cohorts of the same disease. Ideally, different patient cohorts should be analyzed in the same statistical model using the same processing techniques and analysis strategies, or at the very least, standardized reporting should be implemented. Currently, there are several sociological obstacles to such analyses, but projects such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) (The Alzheimer's Disease Neuroimaging Initiative, 2008) may pave the way toward making better use of data. Nevertheless, it is still important that studies provide adequate detail on how they performed their statistical analysis, as well as their preprocessing, in order for the reader to be able to correctly interpret the results (Ridgway et al., 2008).

Summary

In summary, the technique of VBM if implemented correctly is an incredibly powerful and useful tool in the study of neurological disease. It can increase understanding of disease processes, which can be useful both from a scientific point of view and also by providing anatomical information that can be helpful for differential diagnosis of disease. Similar voxel-level statistical techniques can also be applied to other imaging modalities, such as functional MRI and positron emission tomography. It should be stressed, however, that because of the statistical nature of the technique, the power of VBM lies in group analyses. Although it has been applied to single subjects, it has not been optimized or validated for such use. Hence it can provide very important information about regions of atrophy across groups but cannot provide reliable information for single-subject diagnosis. Nevertheless, it is likely to be an important biomarker in future drug trials to assess treatment effects at the group level.

References

Ashburner J, Friston KJ (2000) Voxel-based morphometry—the methods. Neuroimage 11:805–821.

Ashburner J, Friston KJ (2001) Why voxel-based morphometry should be used. Neuroimage 14:1238–1243.

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

Bookstein FL (2001) "Voxel-based morphometry" should not be used with imperfectly registered images. Neuroimage 14:1454–1462.

Davatzikos C (1998) Mapping image data to stereotaxic spaces: applications to brain mapping. Hum Brain Mapp 6:334–338.

Davatzikos C, Genc A, Xu D, Resnick SM (2001) Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. Neuroimage 14:1361–1369.

Davies RR, Scahill VL, Graham A, Williams GB, Graham KS, Hodges JR (2009) Development of an MRI rating scale for multiple brain regions: comparison with volumetrics and with voxel-based morphometry. Neuroradiology. Advance online publication. Retrieved March 24, 2009. doi:10.1007/s00234-009-0521-z.

Friston KJ, Worsley KJ, Frackowiak RS, Mazziotta JC, Evans AC (1993) Assessing the significance of focal activations using their spatial extent. Hum Brain Mapp 1:210–220.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878.

Giuliani NR, Calhoun VD, Pearlson GD, Francis A, Buchanan RW (2005) Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia. Schizophr Res 74:135–147.

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

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

Karas GB, Burton EJ, Rombouts SA, van Schijndel RA, O'Brien JT, Scheltens P, McKeith IG, Williams D, Ballard C, Barkhof F (2003) A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18:895–907.

Keller SS, Roberts N (2008) Voxel-based morphometry of temporal lobe epilepsy: an introduction and review of the literature. Epilepsia 49:741–757.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Prinster A, Quarantelli M, Orefice G, Lanzillo R, Brunetti A, Mollica C, Salvatore E, Morra VB, Coppola G, Vacca G, Alfano B, Salvatore M (2006) Grey matter loss in relapsing-remitting multiple sclerosis: a voxel-based morphometry study. Neuroimage 29:859–867.

Ridgway GR, Henley SM, Rohrer JD, Scahill RI, Warren JD, Fox NC (2008) Ten simple rules for reporting voxel-based morphometry studies. Neuroimage 40:1429–1435.

Rorden C, Bonilha L, Nichols TE (2007) Rank-order versus mean based statistics for neuroimaging. Neuroimage 35:1531–1537.

Rosen HJ, Gorno-Tempini ML, Goldman WP, Perry RJ, Schuff N, Weiner M, Feiwell R, Kramer JH, Miller BL (2002) Patterns of brain atrophy in frontotemporal dementia and semantic dementia. Neurology 58:198–208.

Salmond CH, Ashburner J, Vargha-Khadem F, Connelly A, Gadian DG, Friston KJ (2002) Distributional assumptions in voxel-based morphometry. Neuroimage 17:1027–1030.

Senjem ML, Gunter JL, Shiung MM, Petersen RC, Jack CR Jr (2005) Comparison of different methodological implementations of voxel-based morphometry in neurodegenerative disease. Neuroimage 26:600–608.

Sepulcre J, Sastre-Garriga J, Cercignani M, Ingle GT, Miller DH, Thompson AJ (2006) Regional gray matter atrophy in early primary progressive multiple sclerosis: a voxel-based morphometry study. Arch Neurol 63:1175–1180.

The Alzheimer's Disease Neuroimaging Initiative (2008) MRI methods. J Magn Reson Imaging 27:685–691.

Whitwell JL, Jack CR Jr (2005) Comparisons between Alzheimer disease, frontotemporal lobar degeneration, and normal aging with brain mapping. Top Magn Reson Imaging 16:409–425.

Whitwell JL, Josephs KA (2007) Voxel-based morphometry and its application to movement disorders. Parkinsonism Relat Disord 13 Suppl 3:S406–S416.

Whitwell JL, Josephs KA, Rossor MN, Stevens JM, Revesz T, Holton JL, Al-Sarraj S, Godbolt AK, Fox NC, Warren JD (2005) Magnetic resonance imaging signatures of tissue pathology in frontotemporal dementia. Arch Neurol 62:1402–1408.

Whitwell JL, Jack CR Jr, Boeve BF, Senjem ML, Baker M, Rademakers R, Ivnik RJ, Knopman DS, Wszolek ZK, Petersen RC, Josephs KA (2009) Voxel-based morphometry patterns of atrophy in FTLD with mutations in MAPT or PGRN. Neurology 72:813–820.

Williams LM (2008) Voxel-based morphometry in schizophrenia: implications for neurodevelopmental connectivity models, cognition and affect. Expert Rev Neurother 8:1049–1065.

Wright IC, McGuire PK, Poline JB, Travere JM, Murray RM, Frith CD, Frackowiak RS, Friston KJ (1995) A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 2:244–252.

Ziolko SK, Weissfeld LA, Klunk WE, Mathis CA, Hoge JA, Lopresti BJ, DeKosky ST, Price JC (2006) Evaluation of voxel-based methods for the statistical analysis of PIB PET amyloid imaging studies in Alzheimer's disease. Neuroimage 33:94–102.