NeuroImage **14,** 1238–1243 (2001)
doi:10.1006/nimg.2001.0961, available online at http://www.idealibrary.com on IDEAL®

COMMENTS AND CONTROVERSIES

Why Voxel-Based Morphometry Should Be Used

John Ashburner[1] and Karl J. Friston

The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, United Kingdom

Received May 23, 2001

This article has been written in response to Dr. Fred L. Bookstein's article entitled '"Voxel-Based Morphometry" Should Not Be Used with Imperfectly Registered Images' in this issue of *NeuroImage*. We will address three main issues: (i) Dr. Bookstein appears to have misunderstood the objective of voxel-based morphometry (VBM) and the nature of the continuum we referred to. (ii) We agree with him when he states that findings from VBM can pertain to systematic registration errors during spatial normalization. (iii) His argument about voxelwise tests on smooth data holds in the absence of error variance, but is of no consequence when using actual data. We first review the tenets of VBM, paying particular attention to the relationship between VBM and tensor-based morphometry. The last two sections of this response deal with the specific concerns raised by Dr. Bookstein. © 2001 Academic Press

1. THE OBJECTIVE OF VOXEL-BASED MORPHOMETRY

Structural magnetic resonance (MR) images of brains can differ among subjects in many ways. A useful measure of structural difference among populations is derived from a comparison of the local composition of different brain tissue types (e.g., grey matter, white matter, etc). Voxel-based morphometry (VBM) has been designed to be sensitive to these differences, while discounting positional and other large-scale volumetric differences in gross anatomy.

VBM was originally devised to detect cortical thinning in a way that was not confounded by volume changes of the sort that are characterized by classical volumetric analyses of large brain structures (e.g., the temporal lobe). It does this by removing positional and volume differences (down to a specified spatial scale) through spatial normalization. Differences in grey matter density are then detected by comparing the local intensities of grey matter maps after smoothing. Since its inception some 5 years ago (Wright *et al.*, 1995), VBM has become an established tool in morphometry being used to detect cortical atrophy and differences in slender white matter tracts. The classical perspective on VBM partitions volume changes into two spatial scales: (i) macroscopic volume or shape differences that can be modeled in the spatial normalization procedure and (ii) mesoscopic volume differences that cannot. The latter persist after normalization and are detected after spatial smoothing of grey matter maps (i.e., partitions or segments)—smoothing transforms these volume differences into image intensity differences through the partial volume effect.

More recently, this perspective has changed with the incorporation of an additional step, introduced to compensate for the effect of spatial normalization. When warping a series of images to match a template, it is inevitable that volumetric differences will be introduced into the warped images. For example, if one subject's temporal lobe has half the volume of that of the template, then its volume will be doubled during spatial normalization. This will also result in a doubling of the voxels labeled grey matter. To remove this confound, the spatially normalized grey matter (or other tissue class) is adjusted by multiplying by its relative volume before and after warping. If, warping results in a region doubling its volume, then the correction will halve the intensity of the tissue label. This whole procedure has the effect of preserving the total amount of grey matter signal in the normalized partitions (Goldszal *et al.*, 1998).[2] Classical VBM assumed

[1] To whom correspondence and reprint requests should be addressed at Wellcome Department of Cognitive Neurology, Functional Imaging Laboratory, 12 Queen Square, London WC1N 3BG, UK. Fax: +44 (0)20 7813 1420. E-mail: j.ashburner@fil.ion.ucl.ac.uk.

[2] Note that a uniformly smaller brain will have uniformly lower grey matter intensities after the correction. Any detected differences are therefore less regionally specific, unless some kind of "global" measures are modeled as confounding effects during the statistical analyses. These could pertain to the total amount of grey matter in each brain or, more usefully in many cases, to the intracranial volume of each subject.

1053-8119/01 $35.00

that the warps were so smooth that these volume changes could be ignored. However, advances in normalization techniques now allow for high-resolution warps.

If brain images from different subjects can be warped together exactly,[3] then a complete analysis of the volumetric differences could proceed using only the derived warps. Expansions and contractions occur when an image from one subject is warped to match that of another. These volume changes (hence also the relative volumes of structures) are encoded by the warps. One form of tensor-based morphometry (TBM) involves analyzing these relative volumes (more formally, the Jacobian determinants of the deformation field) in order to identify regions of systematic volumetric difference. The adjustment step mentioned above can be considered from the perspective of TBM, in which volumetric changes derived from the warps are endowed with tissue specificity. By multiplying the relative volumes by the tissue class of interest, volumetric information about other tissue classes is discounted (i.e., there will be no changes attributed to grey matter in purely white matter regions). In other words, the product of grey matter and volume change has two equivalent interpretations: (i) in VBM it represents the proportion of the voxel that is grey matter, having adjusted for the confounding effects of warping the brains, and (ii) it represents the proportion of volume change attributable to grey matter.

By including the multiplication step, a continuum is introduced between the methods of VBM and TBM. At the extreme where little or no warping is used, most of the information pertaining to volumetric differences will be derived from systematic differences in the spatial distribution of the tissue under study. At the other extreme, where registration between images is exact, all the volumetric information will be encoded by the warps (because the normalized partitions will be identical). The product remains sensitive to differences at either extreme and can be regarded as an integration of classical VBM and TBM. This dual perspective is illustrated in Fig. 1, using brain images of a subject suffering from Alzheimer's disease. The data were obtained from the Dementia Research Group (Queen Square, London, UK). This example illustrates progressive volumetric changes of CSF, particularly in the ventricles.

The first column of the figure shows how the later of the two images was processed by classifying the CSF (Ashburner and Friston, 2000) and smoothing it using an isotropic 12-mm FWHM Gaussian kernel. The column in the center shows the processing of the earlier image, which was first rigidly registered with the late image and then the CSF was classified and smoothed. The difference between the smoothed CSF of the two images can be considered as analogous to VBM, where spatial normalization is restricted to a rigid body registration.

The third column shows processing that is analogous to the augmented approach. After rigid registration, the early image is precisely warped to match the late one. The warping method (Ashburner *et al.*, 2000) attempts to estimate exact displacements at every voxel, thus being able to model the relative shapes of the pair of brains. The warped image is then classified, producing an image of the CSF that is very similar to the CSF of the late image. A subtraction of these images would probably not show any meaningful differences. If the segmentation and warping were perfect, then the late CSF image would be identical to the warped early CSF image. To localize CSF volume differences, the volume changes resulting from the warping must also enter into the comparison. To do this, the warped CSF is simply multiplied by the relative volumes estimated from the warps. This means that the procedure preserves the amount of CSF from the original image, while also achieving a good registration. The figure shows an image of relative volumes, where lighter areas indicate a smaller volume in the early image. Following this multiplication, the data are smoothed. The difference between this image and the processed late image shows a picture of volumetric differences based on the warps (cf. TBM). As can be seen from the bottom row of Fig. 1, the pattern of CSF volume differences estimated using the two methods is very similar.

VBM is a simple and pragmatic approach for characterizing small-scale differences that is within the capabilities of most research units. There are many reasons to opt for an approach that involves a less precise registration, but the main one concerns issues of computational requirements. Image registration methods that attempt to estimate an exact match between brain structures are effectively highly nonlinear optimization algorithms with millions of parameters. Fully optimizing a model of this order takes a very long time and is susceptible to local minima, which, in turn, depend upon starting estimates. The alternative, which is what we proposed, is to use a VBM approach with a much simpler warping method, which only attempts to register the brain images "globally."[4] One of the main disadvantages of resorting to a method that uses a less precise image registration is that any re-

[3] Registered so that corresponding brain structures are matched, rather than solving the simpler problem of matching grey matter with grey matter, white with white, and CSF with CSF.

[4] An exact match between any pair of brain images cannot be obtained by a model with only 1000 or so parameters, even if [as Dr. Bookstein suggests (Bookstein, 2001)] the basis set consists of the average template and its changes under large-scale deformation. The use of this "novel" basis set sounds very much like one iteration of the spatial normalization method referred to in the AF paper (Ashburner and Friston, 1999).

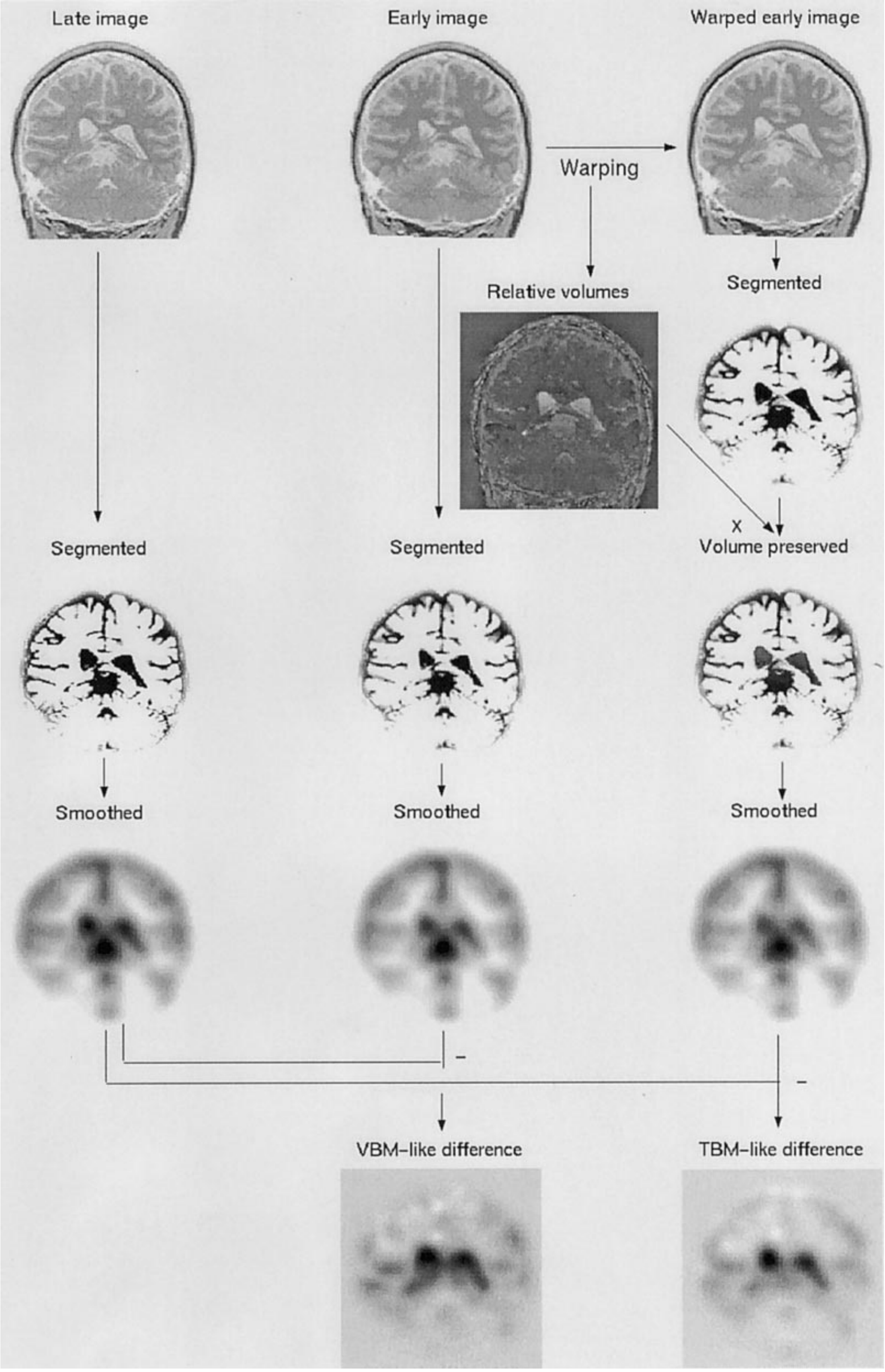

FIG. 1. Illustration of the continuum between VBM and TBM. The center column illustrates a VBM processing stream, whereas the column on the right illustrates a TBM-like stream.

gional volumetric differences cannot be accurately localized.

Dr. Bookstein's article is largely based on a misunderstood premise. When the AF paper (Ashburner and Friston, 2000) refers to the continuum between VBM and TBM, it is alluding to methods sensitive to volumetric differences of brain structures as described above. It is not trying to analyze positional differences. Much of the text and most of the mathematics in Dr. Bookstein's article refer to a "shift functional" which relates to the position of cortical structures, not their volume. His misunderstanding renders his analysis irrelevant to arguments about VBM.

2. SYSTEMATIC REGISTRATION BIASES

Following the preprocessing, which involves spatial normalization, tissue classification, and spatial smoothing, the final step in a VBM analysis is to perform voxelwise statistical tests. The results of these tests are a statistical parametric map (SPM) (Friston *et al.*, 1995a,b) showing significant regional differences among the populations included in the study. Corrections for multiple dependent comparisons are then made using the theory of Gaussian random fields (Friston *et al.*, 1995a,b; Worsley *et al.*, 1996).

Classical statistical tests cannot be used to prove a hypothesis, only to reject a null hypothesis. Any significant differences that are detected could be explained by a number of different causes, which are not disambiguated by the statistical inference. In other words, these tests are not concerned with accepting a specific hypothesis about the cause of a difference, but involve rejecting a null hypothesis with a given certainty. Statistical tests are valid if they produce the correct error rates (i.e., the correct number of false positives). This is what much of the AF paper tried to ascertain. Permutation tests that compared one group with another provided no evidence that the statistical components of SPM99 were invalid for VBM data (smoothed, segmented, and spatially normalized images).[5] The test that transpired not to be valid was based on the spatial extent of excursions. This was because the assumptions about stationarity of smoothness are violated, but this could be remedied in future SPM releases using statistical flattening (Worsley *et al.*, 1999).

Classical statistical tests, as used by VBM, do not protect against type II errors (false-negative results, where real differences are not detected). Dr. Bookstein made several statements about VBM missing real structural differences (Bookstein, 2001). These statements are irrelevant to the discussion as they have nothing to do with the validity of the analysis. We accept that extra sensitivity may be achieved by more accurate warping methods. This argument is the same as for multisubject functional imaging studies, where warping methods with lots of parameters can fractionally increase the sensitivity to activations (Gee *et al.*, 1997).

When the null hypothesis has been rejected, it does not impute a particular explanation for the difference if there are many potential causes that could explain it. VBM manipulates the data so that the ensuing tests are more sensitive to some causes than others. In particular, VBM has been devised to be sensitive to systematic differences in the volumes of grey matter structures, but significant results can also arise for other reasons, some of which are reviewed in this section. Attribution of what may be the cause of any detected difference generally requires a careful characterization of the parameter estimates after the inference has been made.

Dr. Bookstein's main objection to VBM is that it is sensitive to systematic shape differences attributable to misregistration from the spatial normalization step. This is obvious and is one of a large number of potential systematic differences that can arise. For example, a particular subject group may move more in the scanner, so the resulting images contain motion artifact. This motion may interact with the segmentation to produce systematic classification differences. Another group may have systematic differences in the relative intensity of grey matter voxels compared to white matter or may have to be positioned differently in the scanner. All these reasons, plus others, may produce differences that are detectable by VBM. They are all real differences among the data, but may not necessarily be due to reductions in grey matter density.

Of course, a more accurate spatial normalization method would mean that more of the differences can be attributed to volumetric differences in grey matter. However, just because a method has enough freedom to enable it to warp anything to anything else does not always mean that it is a method that is more suited to accurate registration of brain images. More degrees of freedom incur more potential local minima for any warping method that attempts to find the single most probable estimate of the deformations. Furthermore, it becomes nonsensical to attempt to estimate relative shapes of brains beyond a certain spatial scale as the one-to-one mapping between brain regions of different subjects breaks down. For example, many sulci are shared between all brains, but this is not the case for all. Generally the primary sulci, which are formed earliest and tend to be quite deep, are the ones that are the most consistently present. Later developing ones are

[5] As an aside, it is worth noting that no permutation tests have been done that involve comparing single subjects against control groups. Preliminary evidence using nonnormally distributed data of a different type suggests that for these types of comparison, the false-positive rates may differ from those predicted under assumptions of normality. Nonparametric methods (Holmes *et al.*, 1996; Bullmore *et al.*, 1999) may therefore need to be used for these cases.

much more variable. Therefore, parts of some sulci can be objectively matched, whereas others simply cannot.

The basis function approach (Ashburner *et al.*, 1997; Ashburner and Friston, 1999) adopted by SPM99 is far from perfect,[6] but, with the right data, it gives a good global match between brain images. There are cases when structural differences, not directly related to grey matter volumes, can be identified as significant. One obvious example is when one population has larger ventricles. Because the spatial normalization method (Ashburner and Friston, 1999) cannot achieve an exact match, it must change the volume of surrounding tissue when it attempts to make the ventricles of the individual subjects the same size. For example, if the ventricles are enlarged during spatial normalization, then grey matter near them also needs to be enlarged. It is then possible that structural differences pertaining to ventricular volume can show up in a VBM study of grey matter volume. A way of circumventing this would be to base the spatial normalization on only the segmented grey matter. If all the data entering into the statistical analysis were only derived from grey matter, then any significant differences must be due to grey matter.

An analogy can be drawn with some of Bookstein's own work (Bookstein, 1999) on the analysis of shapes of sections through the corpus callosum. How this section is defined influences the resulting cross-sectional shape, as slicing through the brain at a slightly different angle or introducing a small out-of-plane translation will change it. Systematic differences in brain asymmetry among subject groups will produce equally systematic differences among the appearances of the sections. The choice of landmarks used to define the midsaggital section is arbitrary. If an analysis is done with one set of landmarks, one result would be obtained. If done with another, or if the landmarks are selected by a different researcher, then another result may arise. It can be argued that the choice of landmarks is just as arbitrary as the method of spatial normalization used to warp images in VBM studies. How the registration algorithm finds homologous regions is well documented, whereas the rules a human operator uses to identify a homologous point landmark may be more difficult to implement with 100% reliability. Reporting that registration over intersubject variation was done with SPM99 using the default settings is a *precise* description of how homologies are identified. No one has yet defined a *precise* algorithm describing how an investigator manually locates landmarks for morphometric studies.

3. VOXELWISE STATISTICS

There is a near infinite number of ways in which the shapes of brains can differ among populations of subjects. Many thousands or millions of parameters are required to precisely describe the shape of a brain at the resolution of a typical structural MR image. Given, say, 10 schizophrenic and 10 control brain images, there are lots of ways of inventing a measure that would differentiate between the groups. In most cases though, this measure will not provide any distinguishing power in a comparison between further groups of schizophrenics and controls. In other words, the measure would be specific to the subjects included in the study and not generalizable to the populations as a whole. It is therefore not feasible to use methods that try to detect *any* systematic difference. One must be specific about the types of differences that are searched for. Focal differences are often of interest, which is why voxelwise tests are so useful. Volumetric differences of grey or white matter structures are of interest, which are what VBM attempts to emphasize. The data are smoothed spatially prior to performing the voxelwise tests for a number of reasons. The principle reason is to utilize the Matched Filter Theorem in order to sensitize subsequent statistical tests to differences of a particular spatial scale. For example, smoothing the data by 12 mm will make the tests more sensitive to regional differences in structures of about 12-mm extent. Furthermore, smoothing makes the tests better behaved and helps to compensate for inexact spatial normalization.

In the section entitled "Why Not to Use Voxelwise Statistics in Any Event," Dr. Bookstein appears not to approve of doing voxelwise *t* tests on spatially smoothed image data (Bookstein, 2001). The reason given is that signal in a *t* field extends indefinitely, persisting until within-group variance swamps the effect. In real data, residual variance quickly attenuates the *t* statistic to produce the kind of *t* fields that users of methods such as SPM are now familiar with. For this not to be the case, there must be extensive areas of extremely small variance in the residual field or perfect spatial correlations among the errors. As voxels falling outside the brain are normally excluded from VBM analyses, this simply does not happen.

As an example (which is close to a worst-case scenario), consider a simple two-sample *t* test involving a comparison between two groups, each containing 1000 subjects. The standard error of difference (s_D) is 0.0447 times the standard deviation, so a fairly typical standard deviation of 0.01 (see Fig. 2) would correspond to a s_D of 0.000447. The *t* statistic is the difference be-

[6] Brain-warping methods are still in their infancy. Most warping methods involve making MAP-like estimates of deformations, making use of suboptimal guesses for the likelihood and prior probability distributions. A single high-dimensional MAP estimate is not that useful when there are many other possible solutions with similar posterior probability, and estimating expectations over all MAP estimates is not feasible given the high dimensionality of the parameter space.

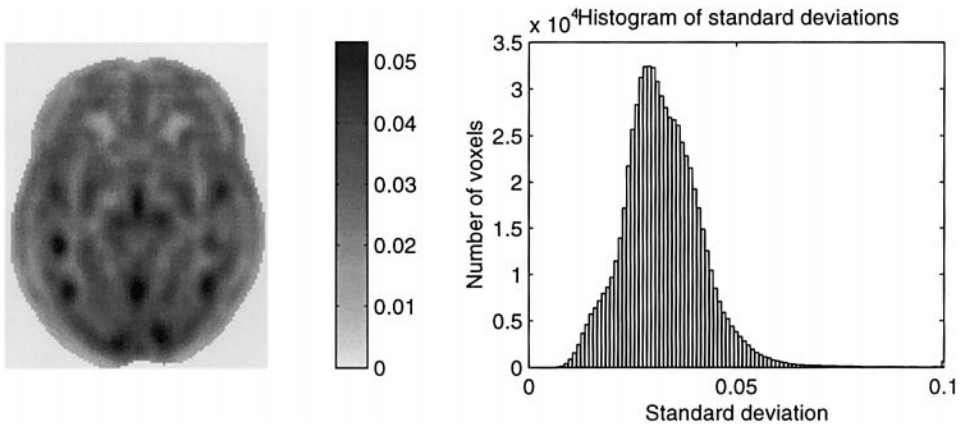

FIG. 2. Typical estimates of standard deviations from a VBM study. On the left is a single plane showing standard deviations. On the right is a histogram showing standard deviations obtained over the whole volume.

tween the groups divided by s_D. If one assumes a very liberal value for α of 0.05, then a voxel will be deemed significant if its t value is greater than 1.6456. If s_D is 0.000447, then the difference must exceed 0.000736 to be considered significant. Using straightforward voxel-based morphometry (without correcting for volumetric differences that arise through spatial normalization), the maximum possible difference between the groups is 1. A 12-mm FWHM Gaussian of amplitude 1 decays to a value of 0.000736, at a distance of 19.4 mm from its center. More realistic group sizes, α value, and magnitude of group differences should demonstrate how specious the objection to Gaussian smoothing of the data really is.

4. SUMMARY

Dr. Bookstein's main criticism made of VBM is that when there are systematic anatomical differences among populations, the method of VBM detects some of them, but not others. In this response we have tried to convey the sorts of differences VBM is interested in and how it is sensitized to them.

ACKNOWLEDGMENTS

Many thanks to the Dementia Research Group (Queen Square, London, UK) for the data shown in Fig. 1. J.A. and K.J.F. are supported by The Wellcome Trust.

REFERENCES

Ashburner, J., Andersson, J., and Friston, K. J. 2000. Image registration using a symmetric prior—In three-dimensions. *Hum. Brain Mapp.* **9**(4): 212–225.

Ashburner, J., and Friston, K. J. 1999. Nonlinear spatial normalization using basis functions. *Hum. Brain Mapp.* **7**(4): 254–266.

Ashburner, J., and Friston, K. J. 2000. Voxel-based morphometry—The methods. *NeuroImage* **11:** 805–821.

Ashburner, J., Neelin, P., Collins, D. L., Evans, A. C., and Friston, K. J. 1997. Incorporating prior knowledge into image registration. *NeuroImage* **6:** 344–352.

Bookstein, F. L. 1999. In *Brain Warping,* Chap. 10, pp. 157–182. Academic Press, San Diego.

Bookstein, F. L. 2001. "Voxel-based morphometry" should not be used with imperfectly registered images. *NeuroImage* **14:** 1454–1462.

Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M. 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transact. Med. Imaging* **18**(1): 32–42.

Friston, K. J., Holmes, A. P., Poline, J.-B., Price, C. J., and Frith, C. D. 1995a. Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage* **4:** 223–235.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C. D., and Frackowiak, R. S. J. 1995b. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2:** 189–210.

Gee, J. C., Alsop, D. C., and Aguirre, G. K. 1997. Effect of spatial normalization on analysis of functional data. In *SPIE Medical Imaging 1997: Image Processing* (K. M. Hanson, Ed.), pp. 312–322.

Goldszal, A. F., Davatzikos, C., Pham, D. L., Yan, M. X. H., Bryan, R. N., and Resnick, S. M. 1998. An image-processing system for qualitative and quantitative volumetric analysis of brain images. *J. Comput. Assist. Tomo.* **22**(5): 827–837.

Holmes, A. P., Blair, R. C., Watson, D. G., Jr., and Ford, I. 1996. Non-parametric analysis of statistic images from functional mapping experiments. *J. Cerebr. Blood Flow Metab.* **16:** 7–22.

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., and Evans, A. C. 1999. Detecting changes in non-isotropic images. *Hum. Brain Mapp.* **8**(2): 98–101.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. 1996. A unified statistical approach for determining significant voxels in images of cerebral activation. *Hum. Brain Mapp.* **4:** 58–73.

Wright, I. C., McGuire, P. K., Poline, J.-B., Travere, J. M., Murray, R. M., Frith, C. D., Frackowiak, R. S. J., and Friston, K. J. 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *NeuroImage* **2:** 244–252.