Title

Voxelwise multivariate analysis of multimodality magnetic resonance imaging

Journal

Human Brain Mapping, 35(3)

ISSN

1065-9471

Authors

Naylor, Melissa G
Cardenas, Valerie A
Tosun, Duygu
et al.

Publication Date

2014-03-01

DOI

10.1002/hbm.22217

Peer reviewed

◆ Human Brain Mapping 35:831–846 (2014) ◆

Voxelwise Multivariate Analysis of Multimodality Magnetic Resonance Imaging

Melissa G. Naylor,[1,2]* Valerie A. Cardenas,[3,4] Duygu Tosun,[3,4] Norbert Schuff,[3,4] Michael Weiner,[3,4] and Armin Schwartzman[1,5]

[1]*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts*
[2]*Pritzker School of Medicine, University of Chicago, Chicago, Illinois*
[3]*Center for Imaging of Neurodegenerative Diseases, Department of Veterans Affairs Medical Center, San Francisco, California*
[4]*Department of Radiology and Biomedical Imaging, University of California, San Francisco, California*
[5]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts*

Abstract: Most brain magnetic resonance imaging (MRI) studies concentrate on a single MRI contrast or modality, frequently structural MRI. By performing an integrated analysis of several modalities, such as structural, perfusion-weighted, and diffusion-weighted MRI, new insights may be attained to better understand the underlying processes of brain diseases. We compare two voxelwise approaches: (1) fitting multiple univariate models, one for each outcome and then adjusting for multiple comparisons among the outcomes and (2) fitting a multivariate model. In both cases, adjustment for multiple comparisons is performed over all voxels jointly to account for the search over the brain. The multivariate model is able to account for the multiple comparisons over outcomes without assuming independence because the covariance structure between modalities is estimated. Simulations show that the multivariate approach is more powerful when the outcomes are correlated and, even when the outcomes are independent, the multivariate approach is just as powerful or more powerful when at least two outcomes are dependent on predictors in the model. However, multiple univariate regressions with Bonferroni correction remain a desirable alternative in some circumstances. To illustrate the power of each approach, we analyze a case control study of Alzheimer's disease, in which data from three MRI modalities are available. *Hum Brain Mapp 35:831–846, 2014.* © 2013 Wiley Periodicals, Inc.

Key words: multivariate analysis; multiple comparisons; multimodality imaging; diffusion tensor imaging; structural magnetic resonance imaging; perfusion weighted magnetic resonance imaging; Alzheimer's disease

Contract grant sponsor: National Institute of Health; Contract grant numbers: T32 MH017119, U01 HL089856, R01 MH081862, R01 MH087590, R21EB013795, and R01 CA157528; Contract grant sponsor: NIH, administered by the Northern California Institute for Research and Education, with resources of the Veterans Affairs Medical Center, San Francisco, California; Contract grant numbers: R03EB008136, P41EB015904, P50AG023501, and P01AG19724.

*Correspondence to: Melissa G. Naylor, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts. E-mail: mgn@uchicago.edu

Received for publication 17 November 2011; Revised 28 August 2012; Accepted 1 October 2012

DOI: 10.1002/hbm.22217

Published online 13 February 2013 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Among the various imaging technologies for studies of the brain, such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography, and single-photon emission CT, MRI stands out by providing a wide range of methods to map various aspects of the brain, including structure, physiology, function, and metabolism. Despite the many MRI modalities available and their often complementary nature, most brain studies have used either a single MRI modality or analyzed each modality separately. Some commonly used MRI modalities include: structural MRI to measure tissue atrophy; diffusion weighted imaging (DWI), which is frequently used to examine the microstructural integrity of white matter (WM); perfusion-weighted MRI, which relates to the arterial blood flow in the brain; functional MRI, which is used to study regional brain activity; and MR spectroscopic imaging, which is used to study cerebral metabolite concentrations. These imaging modalities provide different, but not necessarily independent, information about the brain. By performing an integrated analysis of several modalities simultaneously, new insights may be attained. We refer to the use of two or more MRI methods collectively as multimodality MRI. A joint analysis of multiple modalities not only enables the discovery of effects apparent in two or more modalities simultaneously but also provides information regarding the relationships between modalities.

Researchers have already made some effort to combine information from two different MRI modalities. For instance, to identify regions of concordance and dissociation between structural and perfusion-weighted MRI, without explicitly modeling the correlation between them, Hayasaka et al. (2006) proposed using functions to combine test statistics from different modalities. The significance of the resulting combined statistic was then assessed using permutation tests. A growing number of studies have used various approaches to jointly analyze structural and functional MRI data, including blind source separation methods (Li et al., 2009) and machine learning methods (Lemm et al., 2011). Avants et al. (2008, 2010) used a Hotelling's T^2 test and sparse canonical correlation to jointly analyze structural and diffusion MR images. When reading papers that use multivariate methods such as Hotelling's T^2 to jointly analyze two imaging modalities (Avants et al., 2008), it is natural to ask whether this approach is preferable to doing univariate analyses for each modality and then adjusting for multiple comparisons. If it is, then a more general multivariate regression model can potentially be used to jointly analyze any number of outcomes and account for covariates.

The goal of this article is to answer this question and propose a general statistical methodology that can be used to analyze several MRI modalities simultaneously to increase the statistical power of finding localized characteristics of disease, as well as revealing relationships between the modalities. For this purpose, we assume that the imaging data are a set of coregistered scalar images from a number of subjects, corresponding to various imaging modalities. Since we are doing a voxel-wise analysis, we require spatial registration to be performed on all modalities with the same resolution.

The dataset that motivated this article was a pilot imaging dataset obtained using more than two MRI modalities, including structural MRI, DWI, and perfusion-weighted MRI. This dataset was collected on patients with Alzheimer's disease (AD) and healthy elderly controls to further examine the relationship between macrostructural (structural MRI), microstructural (DWI), and physiological (perfusion-weighted MRI) changes in the brain and the diagnosis of AD. Preliminary exploratory analysis indicated the presence of voxelwise correlations between modalities and this was confirmed by our formal analysis (Data Analysis Results Section).

Although others have documented superiority of multivariate over univariate approaches in general (e.g., Zellner and Huang, 1962), this has rarely been shown in neuroimaging (Young et al., 2010). In neuroimaging specifically, Lazar et al. (2002) and Heller et al. (2007) proposed methods for pooling p-values from various outcomes to test the simultaneous null hypothesis. In this article, we consider whether further gains can be attained by modeling the dependence explicitly via multivariate analysis.

To do this, we first compare the power of multiple voxelwise univariate regressions (one univariate model per modality followed by calculating a combined p-value) with voxelwise multivariate regression (multiple modalities serving as multiple outcomes in a single multivariate model with multiple covariates) in a simulation study. We then perform both types of analysis on experimental multimodality brain MRI data. Although our data is on three modalities of MRI, a joint analysis could similarly be performed on other modalities that provide a single measurement per voxel (or using a voxelwise summary measure from modalities such as fMRI that yield multiple measures per voxel).

METHODS

A common way to examine the relationships between disease and brain alterations is to consider each image voxel individually. For a single modality, each voxel can be thought of as having a vector of image values, with one observed value for each subject. Since we are interested in determining whether there is a relationship between disease status and image values at a specific voxel, we propose a testable hypothesis for each imaging modality:

H_{01}: Disease status does not have an effect on the values observed by a given imaging modality (in a particular voxel).

H_{A1}: Disease status has an effect on the values observed by a given imaging modality (in a particular voxel).

These hypotheses can be tested using univariate linear regression models with disease status as the independent variable and voxel values as the dependent variable, one regression model per modality. The flexibility of such models allows for adjustment of covariates such as age and sex.

However, fitting multiple univariate models to test H_{01} for each modality can lead to false positives due to the increased number of tests. There are several methods for combining p-values in a way that accounts for multiple comparisons. (We review some of these methods in the Testing the Global Null Hypothesis Section.) The resulting combined p-value can be used to reject the global null hypothesis:

H_{02}: Disease status does not have an effect on the values observed by any of the imaging modalities (in a particular voxel).

H_{A2}: Disease status has an effect on the values observed by at least one of the imaging modalities (in a particular voxel).

H_{02} is the intersection of H_{01} for all the modalities of interest. It is also referred to as the intersection null hypothesis, conjunction of null hypotheses, or omnibus null hypothesis (Heller et al., 2007).

Another way to test the global null hypothesis is to test all the modalities at once using a multivariate linear regression model. Each voxel is now thought of as having a corresponding matrix of image values with a column for each modality and a row for each subject. At each voxel, this outcome matrix is treated as the dependent variable in a multivariate multiple regression model. The multivariate model allows us to test the global null hypothesis, H_{02}, in one step. That is, a p-value can be obtained directly from the model to test H_{02}, in contrast to obtaining multiple p-values, each testing H_{01} for a single modality, and then combining them to test H_{02}. The two main advantages of a multivariate analysis, as opposed to separate univariate analyses for each modality, are the potential increase in power attained by taking advantage of correlations between the imaging modalities and the ability to estimate these correlations to better understand the relationships between modalities. In the Univariate Linear Regression and Multivariate Linear Regression Sections, we summarize some known statistical results about univariate and multivariate multiple regression (Mardia et al., 1979), so that we will later be able to discuss in precise terms the similarities and differences between the two approaches.

Univariate Linear Regression

The model

For a single voxel, let $\mathbf{y}$ be the observed vector of the dependent variable on each of n subjects. The dependent variable corresponds to a single imaging modality. In addition, let $\mathbf{X}_{n \times p}$ be a matrix of p covariates on each of the n subjects. At each voxel, we consider a univariate multiple linear regression model of the form

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{u}_i, \tag{1}$$

for the ith modality, where $\mathbf{b}_i$ is a vector of p unknown regression parameters and $\mathbf{u}_i$ is a zero-mean vector of n unobserved Gaussian disturbances, with $\mathrm{Cov}(\mathbf{u}_i) = \sigma_i^2\mathbf{I}$. It is assumed that observations corresponding to different subjects are independent. An unbiased estimator of σ_i^2 is given by $\hat{\sigma}_i^2 = (\mathbf{y}_i - \mathbf{X}\widehat{\mathbf{b}}_i)^T(\mathbf{y}_i - \mathbf{X}\widehat{\mathbf{b}}_i)/(n-p)$. An unbiased estimator of the matrix of regression coefficients $\mathbf{b}$ is given by the maximum likelihood estimator (MLE).

$$\widehat{\mathbf{b}}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_i. \tag{2}$$

Assuming the univariate model (1), H_{01} can be restated as $\beta = 0$, where β, a single element of the vector $\mathbf{b}_i$, is the coefficient of disease. A t-test can easily be used to test H_{01}, as it can be shown that $\widehat{\mathbf{b}}_i \sim N_p(\mathbf{b}_i, \sigma_i^2(\mathbf{X}^T\mathbf{X})^{-1})$.

Testing the global null hypothesis

To test H_{02} using univariate regression, a univariate regression is performed on each outcome and the p-values are then combined into a single joint p-value. Note that this is all still within a single voxel; we will address the multiple testing over voxels in our data analysis in the Data Analysis Results Section. Correction for multiple testing within each voxel is necessary because, if q coefficients for q outcomes are tested, each at level α, then the probability of obtaining false positives at any of the q tests is higher than the desired significance level α.

Lazar et al. (2002) and Heller et al. (2007) have considered several ways to combine multiple tests. An easy way to correct for multiple testing is to use the Bonferroni correction. Typically, the Bonferroni correction is implemented by dividing the threshold of significance by the number of tests, or, equivalently, by multiplying each p-value by the number of tests and using the original significance threshold. Benjamini and Heller (2008) showed that, to use Bonferroni to test the global null hypothesis, H_{02}, one only needs to consider the minimum of the p-values multiplied by the number of tests [see Eq. (3) in their paper]. This is because, if the minimum adjusted p-value does not meet significance, then none of the other adjusted p-values will either. That is, letting $p_1, \ldots, p_q$ denote the p-values for the q outcomes, the p-value for the global null hypothesis, H_{02}, is

$$p_B = \min(qp_1, \ldots, qp_q, 1)$$

For example, if we had the four p-values, 0.01, 0.04, and 0.11, and 0.40, the Bonferroni-corrected p-values would be 0.04, 0.16, 0.44, and 1. To test the global null hypothesis,

H_{02}, we only need to consider whether the minimum of these, 0.04, is less than our significance threshold. If we set the significance threshold at the commonly used 0.05 level, then we can reject the global null hypothesis. This type of correction is guaranteed to control the rate of false positives regardless of the true covariance between the outcomes, which is unknown. Conversely, its generality makes it conservative, especially if q is large. Fortunately, when combining imaging modalities, q will be small.

A major advantage of the Bonferroni method is that it can provide a multiple-comparisons-adjusted p-value for each individual outcome (the p-value multiplied by the number of tests). This is practical for most situations in which the researcher wants to determine which outcomes are associated after the global null hypothesis is rejected. The other methods described here for testing the global null hypothesis only provide a single combined p-value. Even when using the multivariate model, we cannot first test the global null hypothesis and then also test each individual outcome without affecting our type one error.

If the outcomes are independent, then other methods are available to combine the q univariate p-values into a single p-value to test H_{02}. The Stouffer combined p-value is given by

$$p_S = 1 - \Phi\left[\frac{1}{\sqrt{q}}\sum_{i=1}^{q}\Phi^{-1}(1-p_i)\right],$$

where Φ denotes the cumulative distribution function (cdf) of the standard normal distribution. The Fisher combined p-value is given by

$$p_F = 1 - F_{2q}\left[-2\sum_{i=1}^{q}\log(p_i)\right],$$

where F_{2q} denotes the cdf of the χ^2_{2q} distribution. If the outcomes are independent, then the above combined p-values are valid in the sense that their distribution under the null hypothesis is uniform. The disadvantage of these combined p-values is that they are not necessarily valid if the q outcomes are not independent.

Multivariate Linear Regression

The model

As pointed out by Worsley et al. (2004), the simplest approach to analyzing multimodality data is to use a multivariate multiple regression model at each voxel. Similar to the univariate model, the multivariate multiple regression model can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}. \quad (3)$$

This is similar to Eq. (1) except **Y**, **B**, and **U** are now matrices, not vectors. (**X** is the same matrix in both models.) For any given voxel, $\mathbf{Y}_{n\times q}$ is the observed matrix of q outcome variables on each of n subjects. That is,

$$\mathbf{Y} = (\mathbf{y}_1 \quad \mathbf{y}_2 \cdots \mathbf{y}_q),$$

where $\mathbf{y}_i$ for $i = 1, \ldots, q$ are the outcome vectors used in the univariate models. $\mathbf{B}_{p\times q}$ is a matrix of unknown regression parameters and $\mathbf{U}_{n\times q}$ is a zero-mean matrix of unobserved Gaussian errors. It is assumed that observations corresponding to different subjects are independent. Observations corresponding to different outcomes for the same subject may be dependent, but this dependency is assumed to be the same for all subjects. In model (3), this is reflected by the assumption that the rows of $\mathbf{U}_{n\times q}$ have a common covariance $\Sigma_{q\times q}$. Each diagonal element of Σ is the variance of the residuals of a given modality. Each off-diagonal element of Σ is the covariance between the residuals of one modality and the residuals of another modality. It follows that estimating Σ in a multivariate analysis yields estimates of the residual variance for each modality and estimates of the residual covariance between modalities after adjusting for the covariates in the model. An unbiased estimator of the covariance Σ is given by $\hat{\Sigma} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})/(n-p)$. The MLE for **B** is

$$\widehat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (4)$$

This is similar to the univariate case [Eq. (2)], although **B** and **Y** are now matrices. In fact, each of the q columns of $\widehat{\mathbf{B}}$ is equivalent to $\widehat{\mathbf{b}}_i$ in one of the q univariate models. To see this, notice that each column of $\widehat{\mathbf{B}}$ is simply $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ multiplied by the corresponding column in **Y**. Since $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the same in both models and one column of **Y** in the multivariate model is equivalent to a vector $\mathbf{y}_i$ in a univariate model, each column of $\widehat{\mathbf{B}}$ is equivalent to a $\widehat{\mathbf{b}}_i$ in a univariate model. That is, the least-squares (or maximum likelihood) estimation of the mean model parameters results in the same unbiased estimates regardless of whether the univariate or the multivariate model was used. However, the standard errors and p-values obtained will differ.

Testing the global null hypothesis

Scientific questions at each voxel may be formulated in terms of tests of **B**. In their most general form, linear hypotheses may be written as the null hypothesis $\mathbf{CBM} = \mathbf{D}$ for appropriate matrices $\mathbf{C}_{g\times p}$ and $\mathbf{M}_{q\times r}$. For example, suppose FA calculated from DWI is the first of $q = 4$ outcomes and an indicator of disease is the third of $p = 3$ covariates. Then, H_{01} for FA corresponds to $\mathbf{C} = [0\ 0\ 1]$, $\mathbf{M} = [1\ 0\ 0\ 0]^T$, and $\mathbf{D} = 0$, while the global null, H_{02}, corresponds to $\mathbf{C} = [0\ 0\ 1]$, $\mathbf{M} = \mathrm{diag}(1,1,1,1)$, and $\mathbf{D} = [0\ 0\ 0\ 0]$. **M** also allows for testing linear combinations of outcomes, which cannot be done using the univariate model. As an example, if multiple modalities have the

same units, like the multiple outcomes of MRI spectroscopy, then the matrices

$$\mathbf{C} = [0\ 0\ 1], \mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix},$$

and $\mathbf{D} = [0\ 0\ 0]$ test whether disease has the same effect on all the outcomes, as opposed to the effects being different. For any choice of $\mathbf{C}$, $\mathbf{M}$, and $\mathbf{D}$, the likelihood ratio test of the general linear hypothesis $\mathbf{CBM} = \mathbf{D}$ is given by the test statistic

$$\lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}$$

where

$$\mathbf{E} = (n - p)\mathbf{M}^T \hat{\Sigma} \mathbf{M}$$

and

$$\mathbf{H} = \mathbf{M}^T \hat{\mathbf{B}}^T \mathbf{C}^T \left[\mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \right]^{-1} \mathbf{C} \hat{\mathbf{B}} \mathbf{M}.$$

λ has a Wilks' lambda distribution with parameters r, $n - p$, and g. As Wilks' lambda quantiles are difficult to compute, we use Bartlett's χ^2 approximation (Bartlett, 1938). [Note that Rao's F approximation (Rao, 1951) can also be used and generally leads to the same decision.]

Recall that the estimator of $\mathbf{B}$ is the same as we would obtain if we calculated each column of it separately using a standard univariate multiple regression. The test statistics and their distributions used to test the hypothesis H_{02} of no effect on any of multiple modalities, however, are different. In the Simulations section, we use simulations to compare the two approaches.

SIMULATIONS

The goal of the following simulations is to compare the statistical power of the methods described above for testing whether any of the outcomes are related to a covariate of interest. For this purpose, we simulate model (3) for $q = 4$ outcomes and $p = 3$ covariates. The covariates are a vector of ones as the intercept term, a continuous covariate representing standardized age, and a dichotomous covariate representing disease. For each of $n = 20$ subjects, the continuous covariate was produced independently from a standard normal distribution. Half of the subjects were randomly assigned to the disease group and half to the healthy group. For each subject, the model was used to generate data for each voxel in a cube of $10 \times 10 \times 10$ voxels. When generating the data, the regression coefficients were set to zero everywhere except for the central $2 \times 2 \times 2$ voxel cube, where the intercept and the regression coefficient for age were set to 1, while the regression coefficient for the disease was set to either 0 or 2. The number of outcomes generated with a nonzero regression coefficient of disease is a parameter d. We henceforth refer to the d outcomes simulated with nonzero regression coefficients of disease as "truly associated" outcomes since, regardless of the results of any statistical tests subsequently run, they are generated assuming an association between the outcome and disease. Gaussian noise was then added independently at each voxel, independently for each subject. We simulated four different covariance structures for Σ. The diagonal entries (the variances of the outcomes) were set to 1 in all cases. In the first two scenarios, the off-diagonal elements (covariances between the outcomes) are all set to 0 in the independent case, and then 0.5 to produce a strong exchangeable correlations structure. The last two simulated scenarios had more complicated covariance matrices and are discussed later in this section.

In Figure 1, we assume independence and compare the empirical cdfs of uncorrected p-values for testing $B_{31} = 0$. Testing whether $B_{31} = 0$ tests whether disease (the dichotomous third covariate) has an effect on the first outcome (i.e., H_{01}). In each quadrant of the figure, empirical cdfs of the p-values derived from applying multivariate and univariate analyses are shown for a given number of truly associated outcomes, d. Each plot also shows the cdf of p-values under the null hypothesis for reference (note that p-values under the null hypothesis are uniformly distributed; accordingly, the black line is the cdf of a uniform distribution with a range from 0 to 1). A higher curve indicates higher power. The top left panel shows that if one knows a priori which outcome to test, then an uncorrected univariate analysis has better statistical power than a multivariate test of all outcomes. However, if there are two or more dependent outcomes, then the multivariate test is equal or better than the uncorrected univariate test, even before correcting for multiple comparisons. The lower right panel shows that both methods produce uniformly distributed p-values under the null hypothesis of no associated outcomes ($d = 0$). The univariate test is valid here because we are only testing one outcome and therefore do not need to correct for multiple comparisons.

Figure 2 shows a similar comparison under independence when all outcomes are tested and, thus, correction for multiple comparisons is needed. The Bonferroni correction and Fisher's combined p-value perform better than the multivariate method when one or two outcomes (modalities) are dependent on disease status, that is, when $d = 1$ or 2. Stouffer's combined p-value is less powerful than Fisher's in the first panel when only one outcome is truly associated but outperforms Fisher's if there are at least three affected outcomes. The lower right panel shows that all four methods produce uniformly distributed p-values under the null hypothesis of no associated outcomes ($d = 0$). This confirms that the p-values for each of the methods are valid under the assumption of independence.

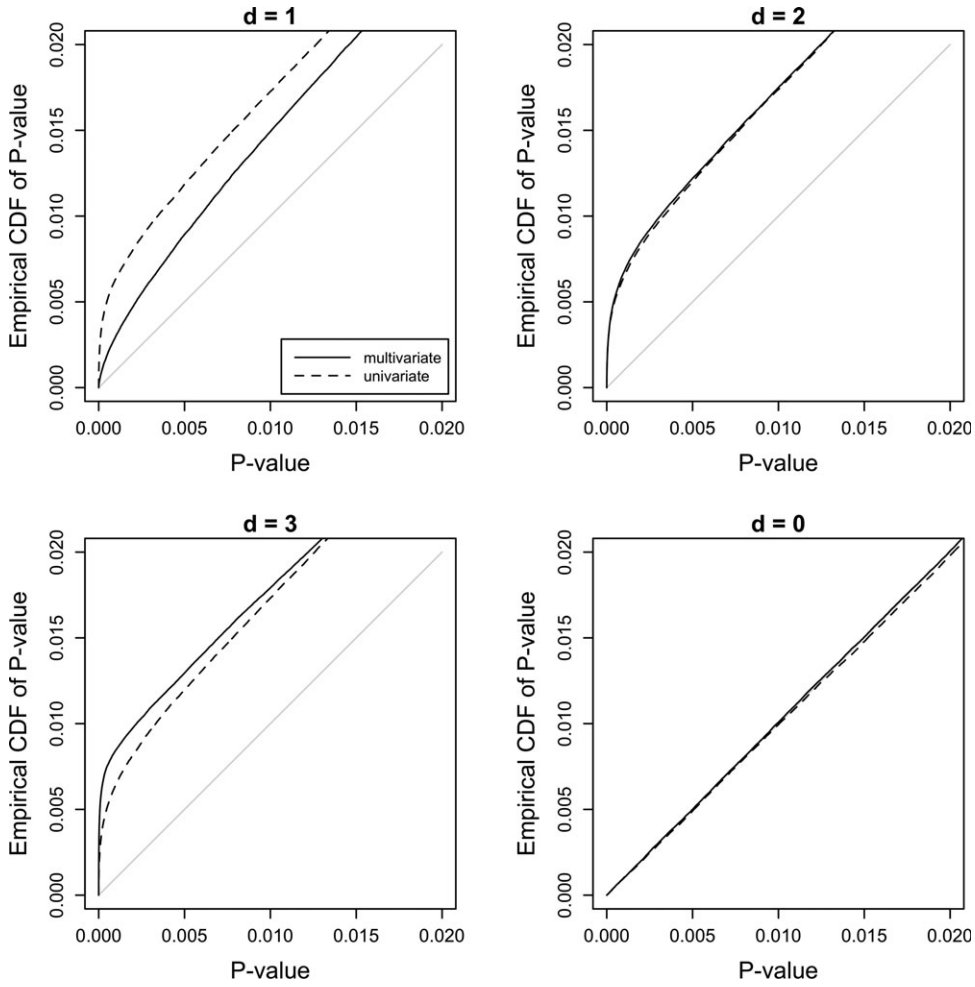

Figure 1.

Empirical cdfs of multivariate and uncorrected univariate *p*-values assuming independent outcomes for: $d = 1$, $d = 2$, $d = 3$, and $d = 0$ truly associated outcomes. The light gray line shows the uniform distribution as a reference.

Figure 3 displays the simulation results assuming pairwise correlations of 0.5 between outcomes. At first glance, it would seem that the Stouffer method performs best, closely followed by Fisher. However, the lower right panel clearly shows that these two methods do not preserve type-one error rate when the tests being combined are correlated; *p*-values generated by the Stouffer and Fisher methods are not valid under dependence. Conversely, the univariate (Bonferroni adjusted) and multivariate methods remain valid. Of the two, the multivariate method performs best, but the Bonferroni correction is not far behind, especially when three outcomes are affected (i.e., $d = 3$). The conservative nature of the Bonferroni correction is evident in the lower right panel where the cdf of Bonferroni corrected *p*-values falls below the 45-degree line.

We were also interested in looking at how the methods perform when the correlation between outcomes is negative. To do this, we simulated data using an extreme correlation matrix estimated using real data (a description of the dataset can be found in the Data Analysis Section). Out of all the gray matter (GM) voxels, we chose the estimated correlation matrix containing the single largest negative element, which was −0.7560350. For the next set of simulations, we used the correlation matrix containing this value:

$$\begin{bmatrix} 1.0000000 & -0.7560350 & -0.2996368 \\ -0.7560350 & 1.0000000 & 0.3001049 \\ -0.2996368 & 0.3001049 & 1.0000000 \end{bmatrix}.$$

The columns (from left to right) and rows (from top to bottom) of the correlation matrix correspond to deformation-based morphometry (DBM) jacobian, fractional anisotropy, and regional cerebral blood flow. Because there were three outcomes in the data used to estimate this correlation matrix, this set of simulations generated three outcomes. Figure 4 displays the simulation results. The

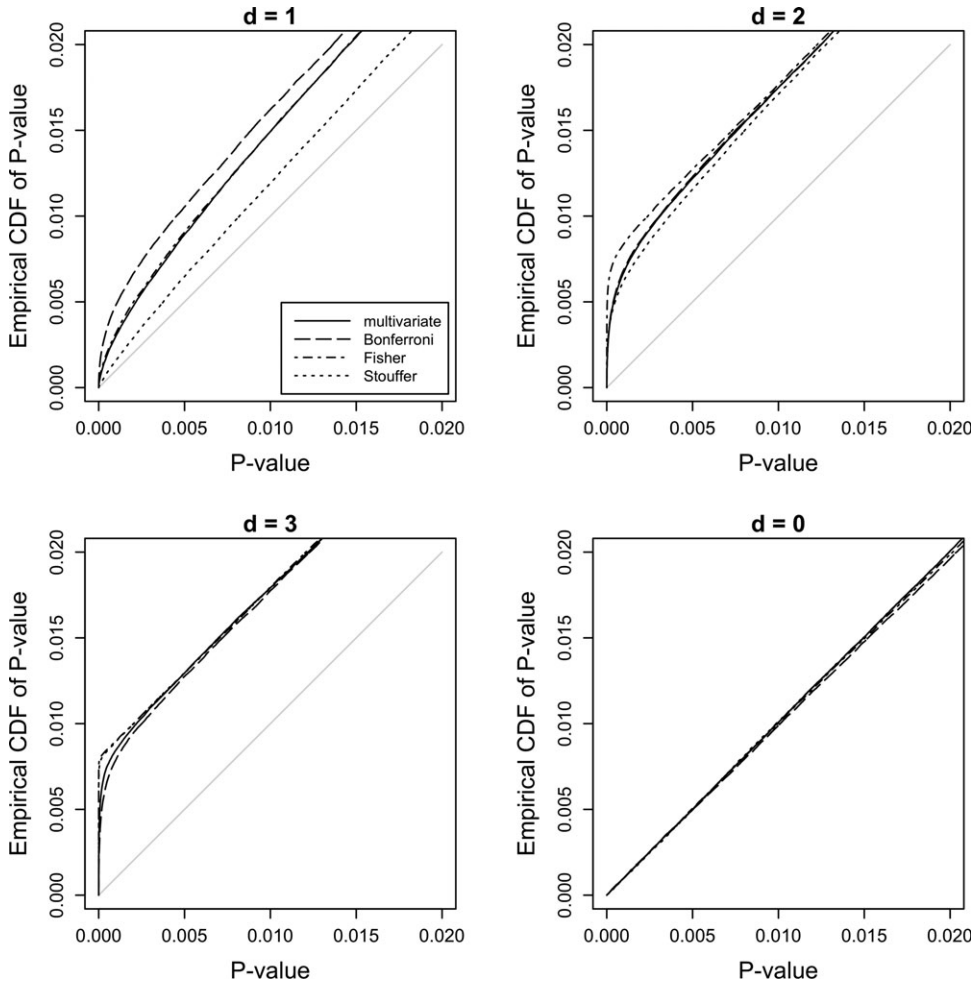

Figure 2.

Empirical cdfs of multivariate and combined *p*-values assuming independent outcomes for: $d = 1$, $d = 2$, $d = 3$, and $d = 0$ truly associated outcomes. The light gray line shows the uniform distribution as a reference.

results are very similar to those seen with exchangeable correlation (Fig. 3), with the most obvious difference being the increase in power when using the multivariate method when $d = 1$. This is likely due to the large (negative) correlation between the first and second outcomes.

To better simulate correlations similar to our real data, we ran simulations in which the 1000 simulated voxels were assigned 1000 correlation matrices randomly selected from the GM voxels in the real dataset. Only GM was sampled as one of the outcomes (regional cerebral blood flow) is not a reliable measure in WM. In the upper left plot of Figure 5, when only one outcome is associated with disease, the Bonferroni correction has higher power, similar to what was seen in the simulations where all outcomes were independent (Fig. 2). This is likely because many of the correlations observed in the real dataset were much smaller than those simulated in Figures 3 and 4. However, when two or three outcomes are truly associated, the multivariate method still has higher power; although the difference in power between the multivariate method and bonferroni correction is not as marked as in Figures 3 and 4.

The simulation results shown in Figures 2–5 are also summarized in Table I, which shows the voxelwise power of each method to detect an association in each of the simulated conditions (i.e., the number of times the affected voxel had a *p*-value less than 0.05 was divided by the total number of simulations). The numbers reiterate the points mentioned above. For example, we see again that Fisher and Stouffer are not valid in the correlated outcomes setting as they have a type one error greater than 0.05 (the last two numbers in the $d = 0$ column).

Across all the simulated scenarios, we see that the multivariate approach has about as much power or more power than Bonferroni correction when there are two or more

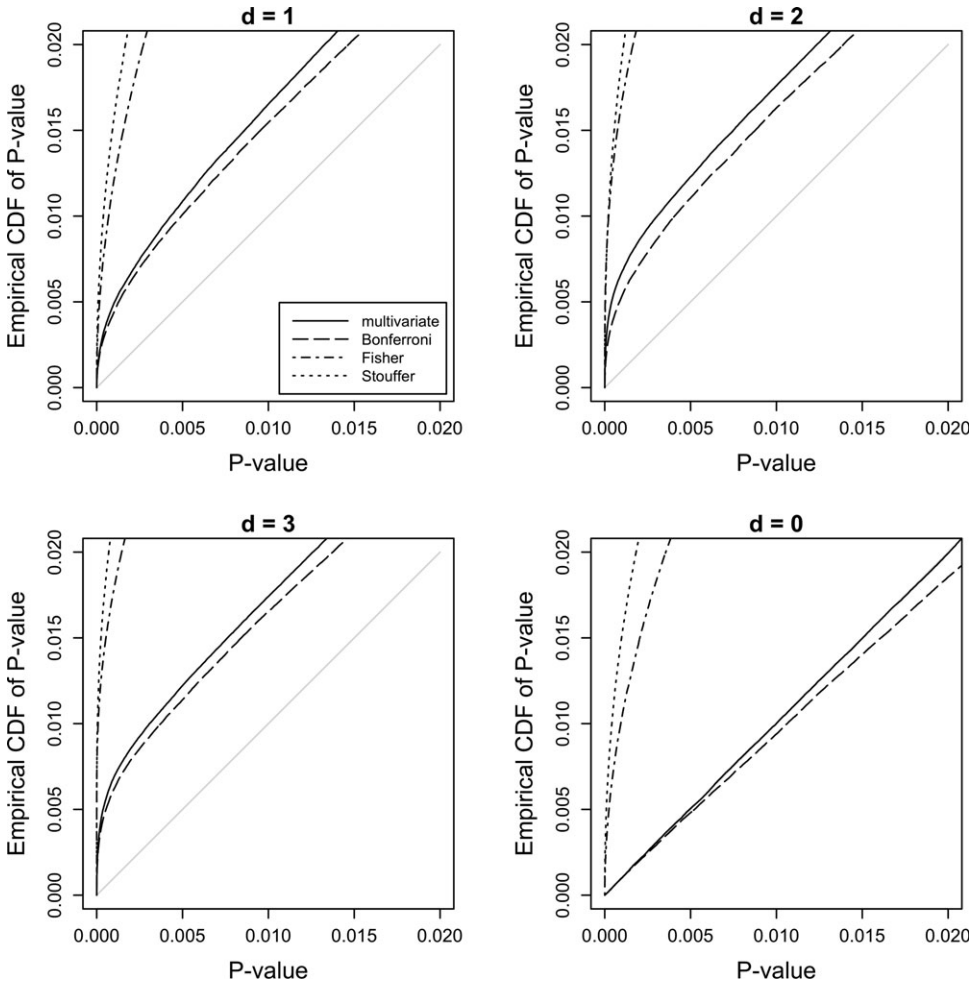

Figure 3.

Empirical cdfs of multivariate and combined *p*-values under exchangeable correlation with $\rho = 0.5$ for: $d = 1$, $d = 2$, $d = 3$, and $d = 0$ truly associated outcomes. The light gray line shows the uniform distribution as a reference.

truly associated outcomes. When only one outcome is truly associated, the multivariate method has higher power when the correlation is exchangeable with $\rho = 0.05$ and when the extreme correlation matrix with large negative correlation was used. However, when the correlations are small or nonexistent as in Figures 5 and 2, respectively, Bonferroni has higher power. In all the simulation results, "power" refers to the probability of rejecting the global null hypothesis given the relationship we assumed when we generated the data. It does not take into account any other testing necessary to determine which outcome(s) are associated once the global null hypothesis is rejected.

DATA ANALYSIS

This cross-sectional MRI study included 84 subjects: 31 with AD (mean age and standard deviation: 65.90 ± 10.22 years; 35.48% females) and 53 elderly healthy control subjects (age: 65.45 ± 9.45 years; 50.94% female). Age, gender, and diagnosis were available for all subjects. Subjects were chosen from a database trying to approximately match age and gender between the AD and control groups. For 19 subjects, Mini-Mental State Examination (MMSE) scores (Folstein et al., 1975) were available; 9 AD patients had an average MMSE score of 21.78 ± 6.06 and 10 controls had an average score of 29.90 ± 0.32.

For each subject, we have brain maps of three quantities, each derived from a separate imaging modality:

1. DBM jacobian (Jacobian): Using data obtained from the sMRI modality, deformation fields relating each brain image to an unbiased atlas brain image were computed. The Jacobian determinant of the deformation fields (i.e., Jac-map), giving the fractional volume contraction or expansion at each voxel evaluated, was

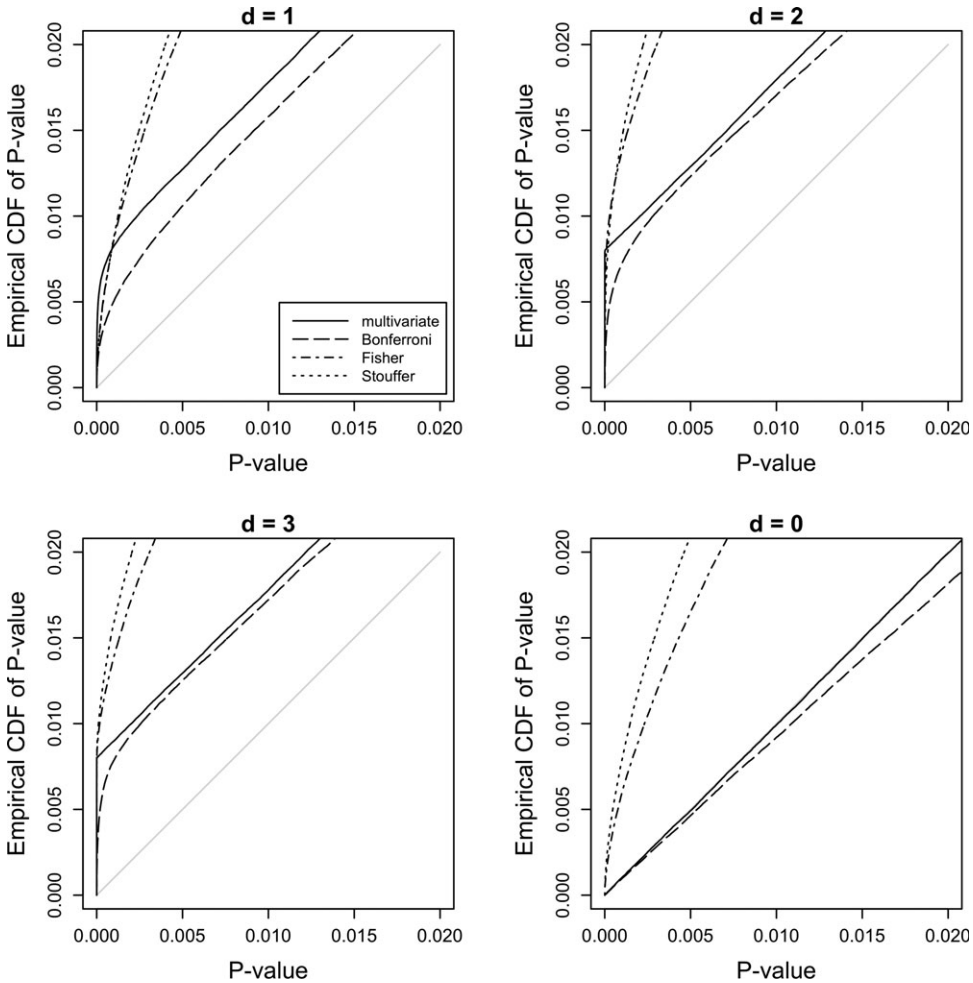

Figure 4.

Empirical cdfs of multivariate and combined *p*-values, under the scenario of an extreme correlation matrix estimated from real data, for: $d = 1$, $d = 2$, $d = 3$, and $d = 0$ truly associated outcomes. The light gray line shows the uniform distribution as a reference.

calculated, providing a map of brain atrophy. Numbers greater (smaller) than one imply local volume expansion (contraction).

2. Fractional anisotropy (FA): FA was computed in the native space of each DTI set (estimated from the DWI modality) after eddy current and geometrical distortion correction using the standard formula for Euclidian FA (Basser and Pierpaoli, 1996). The fractional anisotropy map was then affine coregistered to subjects' T1 images and mapped onto the atlas brain image space using the deformation map from DBM.
3. Regional cerebral blood flood (rCBF): rCBF was based on continuous arterial-spin labeling MRI (Dai et al., 2008) and was recorded as partial volume corrected cerebral blood flow maps, first coregistered to subjects' T1 images and then mapped onto the atlas brain space using DBM deformation field. Average perfusion measurements from the motor cortex region were used for intensity normalization.

Before statistical analysis, spatial smoothing was performed on all images using SPM software package available for Matlab; the contribution of neighboring values to the center pixel was weighted by a Gaussian spatial kernel with a filter width of 10 mm full width at half maximum in the atlas image space. The size of the smoothing kernel matched the size of the effect we sought while accounting for residual errors in the nonlinear spatial normalization. The reference atlas used in calculating DBM was also used to generate probabilistic tissue segmentations for GM, WM, and cerebrospinal fluid.

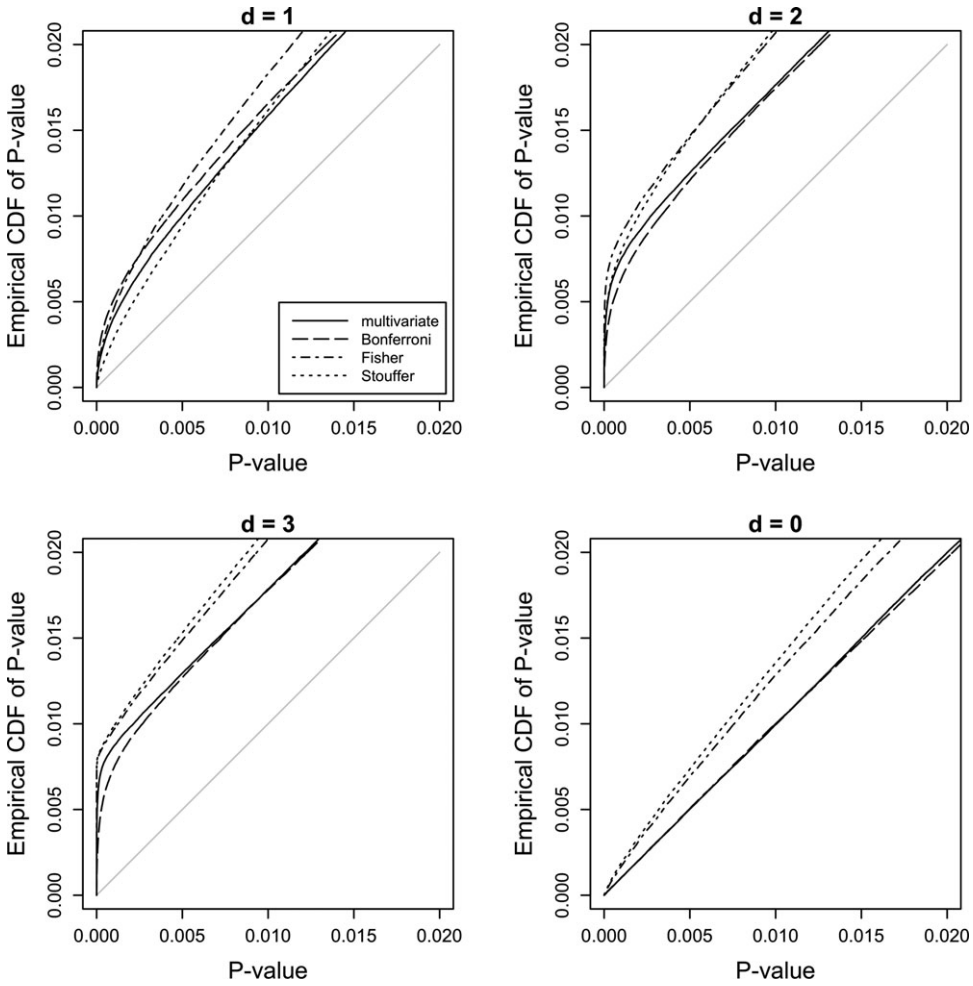

Figure 5.

Empirical cdfs of multivariate and combined *p*-values, where the correlation matrix for each voxel was the estimated correlation matrix from a randomly selected GM voxel in our real data, for: $d = 1$, $d = 2$, $d = 3$, and $d = 0$ truly associated outcomes. The light gray line shows the uniform distribution as a reference.

Subjects

All subjects were recruited from the Memory and Aging Center of the University of California, San Francisco and diagnosed based on information obtained from an extensive clinical history and physical examination. MR images were used to rule out other major neuropathologies such as tumors, strokes, or inflammation but not to diagnose dementia. The subjects were included in the study if they were 55 years old or older and had no history of brain trauma, brain tumor, stroke, epilepsy, alcoholism, psychiatric illness, or other systemic diseases that affect brain function. Subjects received a standard battery of neuropsychological tests including assessment of global cognitive impairment using MMSE and global functional impairment using the Clinical Dementia Rating scale (Morris, 1993). AD patients were diagnosed according to the criteria of the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) (McKhann et al., 1984). All subjects or their legal guardians gave written informed consent before participating in the study, which was approved by the Committees of Human Research at the University of California and the VA Medical Center at San Francisco.

Image Acquisition

All scans were performed on a 4 Tesla (Bruker/Siemens) MRI system with a single housing birdcage transmit and eight-channel receive coil. T1-weighted images were obtained using a 3D volumetric magnetization-prepared rapid acquisition gradient echo sequence with TR/TE/TI

TABLE I. The power of each method to detect an association in a single voxel with a type one error level of 0.05

	$d = 0$	$d = 1$	$d = 2$	$d = 3$
Independent outcomes				
Multivariate	0.0506	0.8623	0.9951	>0.99995
Univariate (Bonferroni)	0.0493	0.9358	0.9961	0.9998
Fisher	0.0491	0.8659	0.9991	>0.99995
Stouffer	0.0509	0.5299	0.9798	>0.99995
Exchangeable Correlation ($\rho = 0.5$)				
Multivariate	0.0475	0.9710	0.9951	0.9973
Univariate (Bonferroni)	0.0420	0.9245	0.9713	0.9861
Fisher	0.0846	0.8321	0.9936	0.9991
Stouffer	0.1000	0.4808	0.9493	0.9995
Extreme correlation matrix from real data containing large negative values				
Multivariate	0.0499	0.9995	>0.99995	>0.99995
Univariate (Bonferroni)	0.0469	0.9604	>0.99995	>0.99995
Fisher	0.0749	0.9313	>0.99995	>0.99995
Stouffer	0.0894	0.6245	0.9981	>0.99995
Each voxel assigned a correlation matrix from a random gray matter voxel in real data				
Multivariate	0.0529	0.9313	0.9981	>0.99995
Univariate (Bonferroni)	0.0501	0.9554	0.9935	0.9991
Fisher	0.0576	0.9121	0.9994	>0.99995
Stouffer	0.0583	0.6371	0.9918	>0.99995

As in Figures 2–5, d is the number of truly associated outcomes; that is, the number of outcomes that were generated as dependent on disease state.

= 2300/3/950 ms, 7-degree flip angle, 1.0 × 1.0 × 1.0 mm^3 resolution, and 157 continuous sagittal slices.

Deformation-Based Morphometry

Each individual skull-stripped and bias field corrected brain image volume was affine registered to a reference brain image to adjust for global differences in brain positioning and scale across individuals. The unbiased average brain used as an atlas was generated from 20 healthy elderly individual brains (i.e., age of 50 to 70 years) using an unbiased atlas formation technique based on large deformations mapping; specifically, a nonlinear inverse-consistent fluid-flow deformation spatially normalized affine registered individual brains to the reference brain (Lorenzen et al., 2005). This set of 20 healthy elderly individuals did not include any controls studied in this work.

Diffusion Weighted Imaging

Diffusion-weighted images were acquired based on a dual-refocused spin-echo echo-planar imaging (EPI) sequence supplemented with twofold parallel imaging acceleration [GeneRalized Auto-calibrating Partially Parallel Acquisitions] (Griswold et al., 2002) to reduce susceptibility distortions. Other imaging parameters of DWI were TR/TE = 6000/77 ms, field of view 256 × 224 cm, matrix size 128 × 112 mm^2, 2 × 2 mm^2 in-plane resolution, and 40 axial slices, each 3 mm thick. Diffusion encoding was accomplished using six, noncollinearily directed magnetic field gradients with $b = 1000$ s/m^2 and normalized to a single EPI without diffusion encoding ($b = 0$). Four diffusion acquisitions were averaged to boost the signal to noise ratio.

Steps involved in preprocessing of DWIs were as follows: Motion and eddy-current distortion artifacts were removed by affine coregistration of each diffusion-weighted image to the unweighted ($b = 0$) image in the diffusion imaging sequence followed by a fluid-flow warping based nonlinear geometry distortion correction to establish anatomical correspondence between structural and diffusion-weighted images. Diffusion tensors were computed using a linear least squares algorithm implemented in TEEM software (http://www.na-mic.org/Wiki/index.php/TeemSummary).

Perfusion-Weighted Imaging

Perfusion-weighted images were acquired using a continuous arterial spin labeling (cASL) sequence (Detre et al., 1992) with a single-shot EPI part and TR/TE = 5200/9 ms timing to map the perfusion signal at 3.75 × 3.75 mm^2 inplane resolution. cASL-MRI consisted of 16 slices, each 5 mm thick and with a 1.2 mm interslice gap. Arterial spin labeling was accomplished with a two-second hyperbolic radiofrequency pulse, followed by an additional one-second postlabeling delay.

Processing of the cASL images involved normalization of the perfusion signal by the arterial water density, differential intensity scaling, and affine alignment followed by the fluid-flow warping based nonlinear geometry

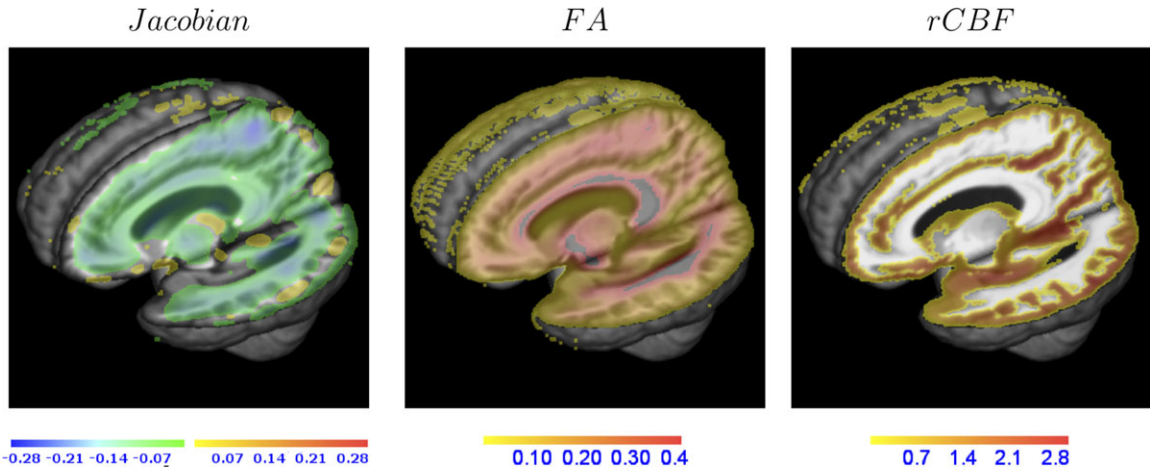

Figure 6.

Average maps for all subjects: average Jacobian map, average FA map, and average rCBF map. The Jacobian is expressed as ratio between the subject-specific volumes and the atlas volume; FA is an index between 0 and 1; rCBF is expressed in institutional units, proportional to blood flow in ml/100mg/min. The colors shown in the color bars are overlayed onto an atlas image with 50% opacity to allow the atlas to remain visible.

distortion correction to establish anatomical correspondence between structural and perfusion images and partial volume correction applied to estimate rCBF in institutional units. rCBF maps were intensity calibrated based on the average rCBF from the sensori-motor cortex.

Data Analysis Results

For illustration, Figure 6 shows the average Jacobian, average FA, and average rCBF maps over all 84 subjects. As expected, we observe high anisotropy in the WM and high perfusion in the GM. The predominately green color of the average Jacobian map merely indicates that, on average, subjects' images had to be contracted to be registered to the atlas image.

As the Jacobian is positive and centered around one, we applied a log transformation to it before analysis. In both the univariate and multivariate models, we adjusted for age and AD diagnosis. No higher-order covariates were used (e.g., age^2, age*disease) as we could not visually detect such relationships when plotting outcomes versus age in a random sample of individual voxels and because we were concerned about the feasibility of estimating an additional three parameters for every covariate added using only 84 subjects.

Figure 7 shows the distribution of voxelwise correlations between the residuals after a given modality has been regressed on covariates including disease status. The histograms show a positive mean correlation between FA and rCBF in the GM, as well as a tendency toward negative correlations between Jacobian and FA in the WM. In each histogram, the mean differed significantly from zero. Figure 8 shows the correlations between residuals of each pair of outcomes. Both positive and negative correlations are observed through out the brain. In the tissue right above the corpus callosum, in which correlation is strongly negative between Jacobian and FA, and between FA and rCBF, but strongly positive between Jacobian and rCBF. Because the residuals are correlated, p-values obtained from univariate and multivariate regression analyses will not necessarily be similar.

We analyzed the data three ways: (A) univariate analysis for a single modality with FDR correction across voxels; (B) univariate analysis for two modalities in WM (Jacobian and FA) or three modalities in GM (Jacobian, FA, and rCBF) with Bonferroni correction for the number of modalities followed by FDR correction for the number of voxels based on the minimum of the three Bonferroni-adjusted p-values; and (C) multivariate analysis of two (in WM) or three (in GM) modalities followed by FDR correction for the number of voxels. The multiple univariate and multivariate analyses were performed as described in Univariate Linear Regression and Multivariate Linear Regression Sections, using maximum likelihood estimation. The FDR correction for testing multiple voxels was performed separately for GM and WM (Genovese et al., 2002). Figure 9 shows maps of t-statistics for the coefficient of disease for each the three modalities, in voxels considered significant by each of the three methods: A, B, or C. The disease covariate is coded 1 for subjects with AD and 0 for controls. As a result, positive t-statistics in Figure 9 indicate an increase in outcome measure for diseased subjects, whereas negative t-statistics indicate a decrease.

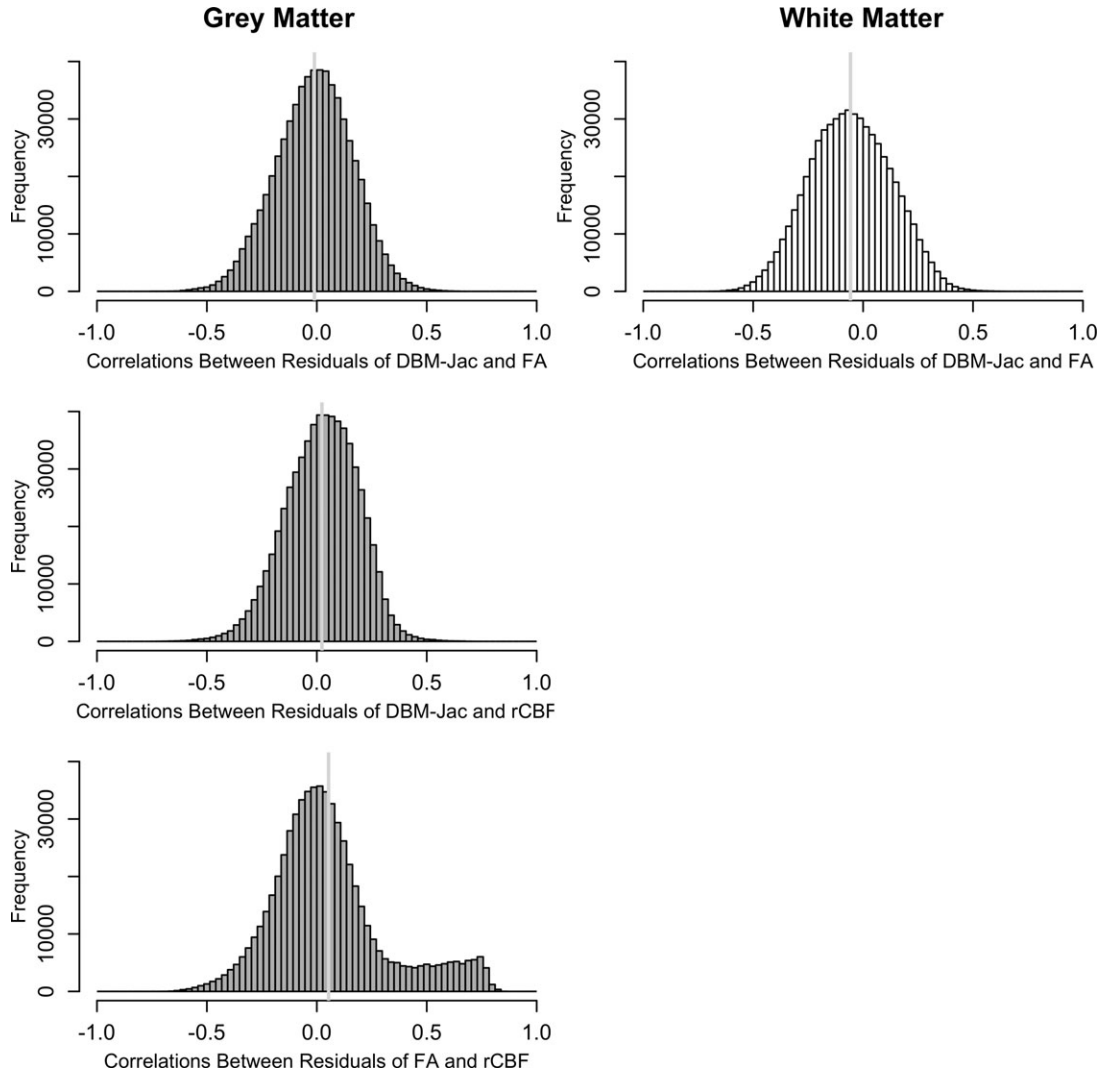

Figure 7.

Histograms of voxelwise correlations between residuals from regressing a given modality on covariates including disease status. Light gray vertical lines indicate the mean. All means are significantly different from zero. A histogram for WM is only shown for correlations between Jacobian and FA because the perfusion signal of WM is not reliable.

When the Jacobian was the dependent variable, univariate analysis revealed significantly smaller tissue volumes in AD patients compared with healthy elderly participants throughout the brain. Though univariate analysis of FA or rCBF revealed several regions of differences between AD and healthy elderly, subjects reductions of FA were observed only in limited regions of the temporal and frontal lobe, and rCBF differences were not observed in the anterior cingulate or temporal lobe. However, when data were analyzed using the multivariate model, it was revealed that atrophy in the frontal lobe, cingulated gyrus, and temporal lobe was accompanied by reductions of FA and differences in rCBF.

DISCUSSION

Maximizing power is important for multiple modality imaging studies in which the total number of statistical tests may be a multiple of the number of modalities and voxels. An adjustment for multiple comparisons must be done to preserve the desired type-one error rate. In this article, we focused on comparing methods available for testing multiple imaging modalities in voxelwise analysis. In particular, we compared multivariate regression analysis to multiple univariate regression models with correction for multiple comparisons. We considered three ways to adjust for multiple comparisons after fitting multiple

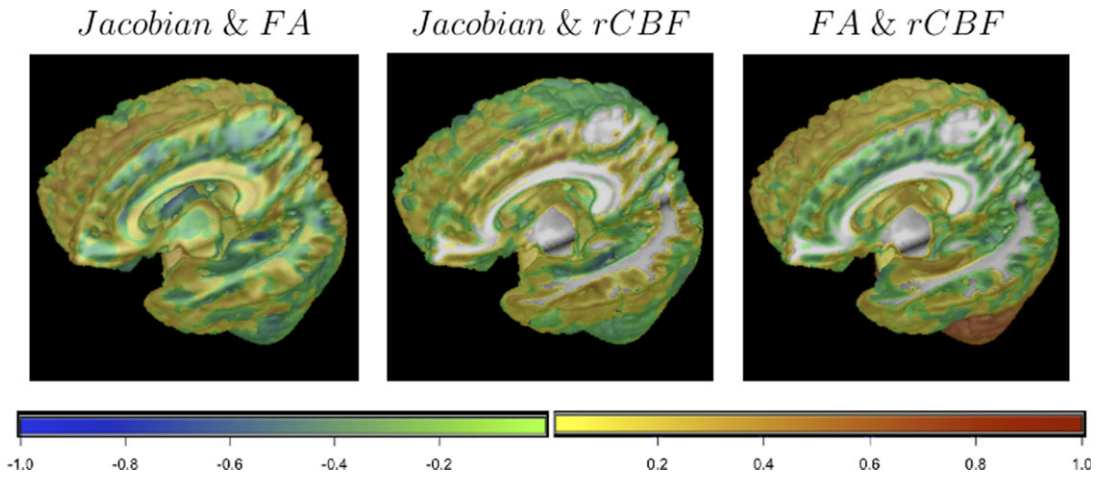

Figure 8.

Correlations between residuals after adjusting for age and disease status are mapped onto the atlas image with an axial cut through left hippocampus and sagittal cut slightly left of midline, revealing the cingulate cortex. Correlations with residuals of rCBF are only shown in GM as rCBF is not a reliable measure in WM.

univariate models: Bonferroni correction, Stouffer's combined *p*-value, and Fisher's combined *p*-value. Our simulations confirmed that as Stouffer's and Fisher's combined *p*-values assume independence between tests, they do not produce valid *p*-values when the tests are correlated. Since it cannot be assumed that multiple imaging modality measures on the same voxel are independent, these methods are not appropriate. In comparing the multivariate model to multiple univariate models followed by Bonferroni correction, we found that the multivariate model has higher power to reject the global null hypothesis if more than two outcomes were associated with covariate of interest or if the outcomes were highly correlated. Conversely, if only one or two outcomes are associated with the covariate of interest and the outcomes were independent, fitting multiple univariate models and applying Bonferroni correction is more powerful.

Beyond simply maximizing power to reject the global null hypothesis, there are other issues, such as the hypothesis of interest and computational efficiency, that should be taken into account when choosing between multivariate and multiple univariate analysis. For instance, if estimating the correlations is of inherent interest, then certainly the multivariate model is preferable. Likewise, if the main objective is to find voxels where more than one outcome is associated, then the multivariate model provides the most powerful way to do this. On the contrary, if easily obtaining a *p*-value for each outcome that accounts for multiple comparisons is highly desirable, multiple univariate analyses with Bonferroni correction are a better option. Using Bonferroni correction may also be more computationally thrifty. Fitting a multivariate model at each voxel can be computationally intensive when the number of voxels is large. In our case, the multivariate fitting was coded in C (http://www.nitrc.org/projects/valmap), as other implementation attempts in Matlab or R were excessively time-consuming. The Bonferroni combined *p*-value was much simpler to implement. In the simulations and data analysis, we saw that the Bonferroni combined *p*-value does not perform as well as the multivariate model, but it did not lag far behind, yielding similar significance regions in Figure 9. In addition to power, the hypothesis of interest, ease of implementation, and other issues are important to consider when choosing a method of analysis.

In our analysis of experimental data, we demonstrated both multivariate analysis and univariate analysis with Bonferroni correction. We sought to boost power and gain anatomical insight into imaging of neurodegenerative diseases by jointly analyzing structural, perfusion-weighted, and diffusion-weighted MRI measures. Multivariate analysis allowed for detection of the effect of disease on each modality and also helped elucidate relationships between the various modalities with respect to the disease process. In Figure 9, voxels with statistics large in magnitude (shown as red or blue) in more than one modality indicate simultaneous effects of the disease on these modalities. For instance, the Jacobian maps show enlargement of the ventricles, while the FA maps show a simultaneous reduction in FA in the corpus callosum. It is known that the regression coefficients in a multivariate model are the same as those in the separate univariate models. Therefore, simultaneous effects such as the above could also be inferred from separate univariate analyses. The main advantage of the multivariate approach is to be able to detect larger areas of significance, thus detecting brain changes that may not be visible in the univariate analyses.

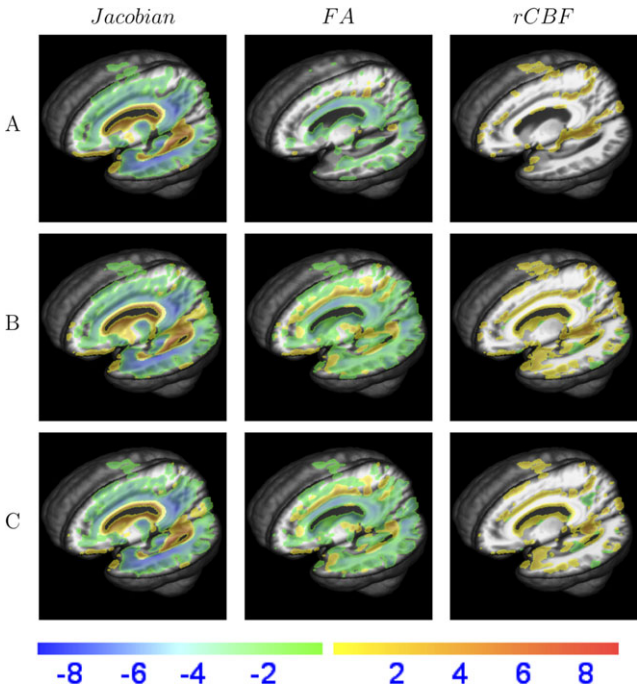

Figure 9.

Significant associations with disease status after adjusting for age are shown with an axial cut through left hippocampus and sagittal cut slightly left of midline, revealing the cingulate cortex. Perfusion was only modeled in the GM. In each image, we adjusted for testing multiple voxels by applying FDR correction separately for GM and WM. A voxel is considered significant if the corresponding q-value < 0.05. Colors indicate the value of t-statistics for the effect of disease status on the modality indicated at the top of the column. Note that the disease covariate is coded one for subjects with Alzheimers disease and zero for controls. Row A shows the voxels significant after univariate analysis. Row B shows the voxels significant when Bonferroni correction is used to adjust for testing two or three outcomes separately (two in WM; three in GM). The minimum of the three Bonferroni-adjusted *p*-values for each voxel was used for FDR correction. Row C shows the voxels significant based on Wilks' lambda from a multivariate analysis.

In this article, we considered multivariate modeling of multiple outcomes for the same voxels, but not joint modeling of multiple outcomes at different voxels. The FDR correction performed over voxels was marginal and did not consider the dependence between voxels. It is likely that detection power could be improved and more information could be gained by joint multivariate modeling between voxels or by applying other methods that account for spatial correlations between voxels (Bowman, 2005; Derado et al., 2010). For example, random field theory, which accounts for spatial correlations between voxels, could be used instead of FDR to control for multiple testing over voxels (Worsley et al., 2004). These methods are also attractive from the perspective of the neural connectivity of the brain. The study of these more complex methods is left for future work.

Although we have only demonstrated the analysis of three modalities of MRI, the statistical concepts are general, and are, in principle, applicable to other imaging technologies. However, some modalities, such as fMRI which produces an entire time series of outcomes for each voxel, may require adaptions of the methods presented here. Importantly, the interpretation of the correlations will depend on the modalities analyzed.

ACKNOWLEDGMENT

The authors thank Dr. PT Fletcher, University of Utah, for providing software for geometrical distortion corrections.

REFERENCES

Avants B, Duda JT, Kim J, Zhang H, Pluta J, Gee JC, Whyte J (2008): Multivariate analysis of structural and diffusion imaging in traumatic brain injury. Acad Radiol 15:1360–1375.

Avants BB, Cook PA, Ungar L, Gee JC, Grossman M (2010): Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. Neuroimage 50:1004–1016.

Bartlett M (1938): Further aspects of the theory of multiple regression. Proc Cambridge Philos Soc 34:33–40.

Basser PJ, Pierpaoli C (1996): Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. J Magn Reson B 111:209–219.

Benjamini Y, Heller R (2008): Screening for partial conjunction hypotheses. Biometrics 64:1215–1222.

Bowman FD (2005): Spatio-temporal modeling of localized brain activity. Biostatistics 6:558–575.

Dai W, Garcia D, de Bazelaire C, Alsop DC (2008): Continuous flow-driven inversion for arterial spin labeling using pulsed radio frequency and gradient fields. Magn Reson Med 60:1488–1497.

Derado G, Bowman FD, Kilts CD (2010): Modeling the spatial and temporal dependence in FMRI data. Biometrics 66:949–957.

Detre JA, Leigh JS, Williams DS, Koretsky AP (1992): Perfusion imaging. Magn Reson Med 23:37–45.

Folstein MF, Folstein SE, McHugh PR (1975): "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12:189–198.

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

Griswold MA, Jakob PM, Heidemann RM, Nittka M, Jellus V, Wang J, Kiefer B, Haase A (2002): Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 47:1202–1210.

Hayasaka S, Du A, Duarte A, Kornak J, Jahng G, Weiner M, Schuff N (2006): A non-parametric approach for co-analysis of multi-modal brain imaging data: Application to alzheimer's disease. Neuroimage 30:768–779.

Heller R, Golland Y, Malach R, Benjamini Y (2007): Conjunction group analysis: An alternative to mixed/random effect analysis. Neuroimage 37:1178–1185.

Lazar N, Luna B, Sweeney J, Eddy W (2002): Combining brains: A survey of methods for statistical pooling of information. Neuroimage 16:538–550.

Lemm S, Blankertz B, Dickhaus T, Müller KR (2011): Introduction to machine learning for brain imaging. Neuroimage 56:387–399.

Li YO, Adal T, Wang W, Calhoun VD (2009): Joint blind source separation by multi-set canonical correlation analysis. IEEE Trans Signal Process 57:3918–3929.

Lorenzen P, Davis B, Joshi S (2005): Unbiased atlas formation via large deformations metric mapping. Med Image Comput Comput Assist Interv 8:411–418.

Mardia K, Kent J, Bibby J (1979): Multivariate Analysis. New York: Academic Press.

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984): Clinical diagnosis of alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on alzheimer's disease. Neurology 34:939–944.

Morris JC (1993): The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology 43:2412–2414.

Rao CR (1951): An asymptotic expansion of the distribution of Wilks' criterion. Bull Int Stat Inst 33:177–180.

Worsley K, Taylor J, Tomaiuolo F, Lerch J (2004): Unified univariate and multivariate random field theory. Neuroimage 23:S189–S195.

Young K, Govind V, Sharma K, Studholme C, Maudsley AA, Schuff N (2010): Multivariate statistical mapping of spectroscopic imaging data. Magn Reson Med 63:20–24.

Zellner A, Huang D (1962): Further properties of efficient estimators for seemingly unrelated regression equations. Int Econ Rev 3:300–313.