Accounting for Spatial Dependence in the Analysis of SPECT Brain Imaging Data

Jeffrey S. Spence, Patrick S. Carmack, Richard F. Gunst,
William R. Schucany, Wayne A. Woodward, Robert W. Haley *

*Jeffrey S. Spence and Patrick S. Carmack are Assistant Professors of Biostatistics and Robert W. Haley is Professor and Head, Epidemiology Division, Departments of Internal Medicine and Clinical Sciences, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-8874, USA. Richard F. Gunst is Professor and Chair, and William R. Schucany and Wayne A. Woodward are Professors, Department of Statistical Science, Southern Methodist University, P.O. Box 750332, Dallas, TX 75275-0332, USA. This study was supported by the U.S. Army Medical Research and Materiel Command numbers DAMD17-97-2-7025 and DAMD17-01-1-0741 through a consortium agreement with the University of Texas Southwestern Medical Center at Dallas, U.S. Public Health Service grant M01-RR00633, and by a grant from the Perot Foundation, Dallas, Texas. The content of this paper does not necessarily reflect the position or the policy of the U.S. government, and no official endorsement should be inferred. Email: rgunst@smu.edu.

ABSTRACT

The size and complexity of brain imaging databases confront statistical analysts with a variety of issues when assessing brain activation differences between groups of subjects. Detecting small group differences in activation is compounded by the need to analyze hundreds of thousands of spatially correlated measurements per image. These analyses are especially problematic when, as is typical, the number of subjects in each group is small. In this article, a comprehensive analysis of single-photon emission computed tomography (SPECT) brain images demonstrates that spatial modeling can increase the sensitivity of group comparisons. The key statistical approach for increasing the sensitivity of group comparisons is the spatial modeling of intervoxel correlations. Correlations among normalized SPECT counts in $2 \times 2 \times 2$ mm^3 voxels are shown to be very large in neighboring voxels and decrease in magnitude until they become negligible among those approximately 5-7 voxels (10–14 mm) apart. Exploiting this correlation structure, blocks of contiguous voxels are defined within each of several structures within the deep brain so that the geometric centers of the blocks are no closer than the range at which voxel counts can be considered uncorrelated. Using kriging methods, block averages and their prediction variances are calculated. For each structure of interest the block averages within the structure are weighted by their prediction variances, producing a structure average for each subject. The subject averages and their prediction variances are used in a linear model to compare group effects. This analysis is shown to be more sensitive to group mean differences than the voxel-by-voxel analysis commonly used by medical researchers. The procedures are applied to comparisons of SPECT brain-imaging data from four groups of subjects, three of which have variants of the 1991 Gulf War syndrome and one of which is a control group. Commonly used voxel-by-voxel group comparisons do not identify any brain structures that are significantly different for the syndrome and control groups in the analysis of cholinergic response to a physostigmine drug challenge. Spatial modeling and analyses of these data do identify regions of the deep brain that exhibit statistically significant group differences. These results are consistent with medical evidence that these structures might have been affected by Gulf War chemical exposures.

KEY WORDS: Gulf war syndrome, Kriging, Multiple comparisons, Semivariogram, Single-photon emission computed tomography

1. INTRODUCTION

United Nations military personnel returning from the 1991 Persian Gulf War reported post-war changes in mental function and physical performance that did not seem reasonably attributable to known maladies such as infectious diseases or post-traumatic stress disorder. Haley, Kurt, and Hom (1997a) analyzed medical records, wartime exposures, responses from a survey instrument of illness symptoms, and psychological profiles of 249 U.S. Naval reservists of the 24^{th} Reserve Naval Mobile Construction Battalion (seabees) who were stationed in the Gulf War theater of operations. They characterized three primary medical syndromes that were shared by approximately one-third of the 175 seabees who complained of serious health problems that the seabees believed resulted from service in the Gulf War. Kang et al. (2002) confirmed the syndromes in a stratified sample of all the military personnel who were deployed in the 1991 Gulf War. In a clinical study of 26 affected Gulf War veterans and 20 matched controls, Haley et al. (1997b) and Hom, Haley, and Kurt (1997) concluded that the observed abnormalities in brain functioning were consistent with subcortical and brainstem dysfunction. In an epidemiological study of self-reported risk factors of the 249 seabees, Haley and Kurt (1997c) found that the three primary syndromes were consistent with chemical exposures but not with other proposed risk factors such as oil well fires or combat stress. Postulating that the syndromes were due to deep-brain irregularities or malfunctions, a subsequent study of 39 syndrome and control seabees using single-photon emission computed tomography (SPECT) brain imaging was undertaken. Commonly used analyses of brain images at individual brain locations were not able to detect any group differences in brain activity. This article reports the successful detection of syndrome and control group differences. The application of spatial statistical modeling was critical to the detection of the group differences.

Impaired cognition (Syndrome 1) is characterized by distractibility, difficulty remembering, depression, insomnia, confused thought processes, and migraine headaches. Confusion ataxia (Syndrome 2) is characterized by difficulty with thinking and reasoning, confusion with where one is or what one is doing, disorientation, balance problems, sensations of rooms spinning, stumbling, physician diagnosis of post-traumatic stress disorder, liver disease, or sexual impotence. Central pain (Syndrome 3) is characterized by joint and muscle pain, muscle fatigue, tingling, and numbness. Haley et al. (1997c, 1999, 2000) attributed these syndromes to neurological damage, possibly from exposure to chemical nerve agents, anti-nerve gas medication, pesticides, or high concentra-

tions of insect repellants in genetically predisposed individuals. They postulated that exposure to these agents might have caused damage in the basal ganglia or the brainstem, which are gray-matter structures deep within the brain (see Figure 1).

[Insert Figure 1]

In an attempt to identify areas of the deep brain that might have been affected by exposures in the Gulf War, an experiment was conducted with the voluntary participation of 5 Syndrome 1 subjects, 12 Syndrome 2 subjects, 5 Syndrome 3 subjects, and 17 control subjects from the original study of 249 seabees. The small numbers of subjects in each group reflect the costs (medical personnel, housing, medications, and tomography) incurred in performing multiple brain scans on each of these subjects. The larger relative number of subjects in the Syndrome 2 group is due to previous neurological studies that indicated this group was the most severely affected (Haley 2000, 2002). The control subjects were closely matched by age, gender, and education to the Syndrome 2 subjects. Nine of the controls served in the Gulf War theater of operations but did not subsequently report any adverse health effects. Eight of the controls remained in the United States. Since no substantive differences were detected between the two groups of control subjects, they were combined for the analyses reported below.

In December 1997 – June 1998, SPECT scans of brain activity were obtained on each subject during a baseline session and 2-5 days later in a treatment session. The goal of the SPECT imaging was to determine the extent of brain activity through measurements of the rate of cerebral blood flow. A radioactive tracer was introduced into each subject's blood through a rapid intravenous injection. Differential blood flow in the brain causes more radiation to be emitted in active brain areas where the rate of blood flow is relatively high and less to be emitted in areas where blood flow is reduced by inactivity, normal inhibition, or impairment. In the baseline session, the patients were administered the radioactive tracer after an intravenous infusion of a placebo, saline. In the treatment session, physostigmine was added to the saline infusion. Physostigmine elevates acetylcholine, which typically inhibits brain activity in normally functioning gray-matter regions of the brain. It was anticipated that healthy brain regions for both syndrome and control groups would exhibit inhibited gray-matter brain activity in the treatment session. While it was suspected that dysfunctional gray-matter brain regions in the syndrome subjects that were damaged by Gulf

War exposures would show little or no change between the baseline and treatment sessions, there was no conclusive information from previous clinical investigations about the nature of the effects of physostigmine on dysfunctional brain regions.

SPECT intensity counts are available for approximately 200,000 $2 \times 2 \times 2$ mm^3 volume elements, or *voxels*, in each of the 39 subject brains at each measurement session. SPECT measurements are highly variable among individuals, between sessions for each individual, and among voxels within a structure at one session for one individual. Because of the large number of measurements for each three-dimensional image, the high variability of measurements in each image, and the small number of subjects in each group, informative statistical analyses of the SPECT database require that a number of critical data processing issues be addressed prior to any comparisons of brain measurements. These preprocessing issues are now briefly summarized.

2. BRAIN IMAGE PREPROCESSING

Mapping brain locations across subjects and sessions is complicated by problems with alignment of subjects in the radiation-detecting scanner, head movements as measurements are being taken, and differences in brain sizes and shapes. Several preprocessing steps were critically examined for use with the Gulf War multi-subject, multi-session SPECT measurements.

First, registration of the baseline and treatment SPECT scans to a magnetic resonance image (MRI), a highly accurate anatomical scan of an individual's brain, was attempted. Statistical Parametric Mapping (SPM) (Friston 2004) is a widely used software program that enables both the preprocessing of images and comparative statistical analyses to be performed. SPM mapping of the SPECT scans to the MRI images introduced additional variability without demonstrably improving the intra-individual voxel registration. This is possibly because the anatomical scans were taken two years prior to the baseline and treatment SPECT scans. Hence, no mapping to individual MRI scans was used.

Second, the SPM spatial normalization algorithm was used to map all brain images to the standard three-dimensional space of an SPM SPECT template, coordinates of an average brain template (MNI152) compiled from 152 highly accurate anatomical scans taken by the Montreal Neurological Institute. This is a necessary mapping for group comparisons.

Third, kernel smoothing methods, often used to better reflect brain activity in clusters of voxels and to accommodate the likelihood that voxel locations are not exactly the same in each registered scan, were investigated. Kernel smoothing was discarded because simulations indicated that these spatial averages might mask differences in intensities within specified deep-brain gray-matter structures of interest. The masking can occur because meaningful differences in signal intensity within small structures are sometimes averaged with smaller inhomogeneous differences in intensity from neighboring gray-matter structures, white-matter structures, or ventricles filled with cerebrospinal fluid.

A fourth preprocessing step is necessary because of the known rapid decay of radioactive tracers. Even with the greatest of experimental care, intra- and inter-individual differences in tracer decay rates occur because of uncontrollable session-to-session variation in the exact amount and timing of tracer injections, equipment setup and calibration, and subject differences in metabolism and brain activity. Consequently, intensity counts for each individual and each session must be scaled to a comparable magnitude, a procedure called count normalization. Spence et al. (2006) conducted an extensive investigation of the count normalization procedures in common use. They recommended normalizing voxel counts by the median count from an easily identified, large white-matter region, the centrum semiovale. All Gulf War SPECT voxel intensities were normalized by dividing the individual voxel counts in a scan by the centrum semiovale median white-matter count and then scaled so that the median normalized intensity count in each scan's centrum semiovale equals 100.

Finally, since Haley et al. (1999, 2000) postulated that syndrome effects were due to dysfunction of one or more structures within the deep brain, individual structures or regions of interest (ROIs) within the deep brain were identified from the MNI standard coordinate system through the use of the Talairach atlas (Talairach and Tournoux 1988). The Talairach daemon (Lancaster et al. 2000) was used to identify intracranial structures by specifying coordinates of the voxel locations. The Carmack et al. (2004) affine transformation was then used to accurately map the Talairach coordinates of structures to the MNI152 standard coordinates. Care was taken to minimize partial volume effects due to the selection of voxels for an ROI that contains portions of neighboring structures. A total of 16 ROI databases were created from the right and left halves of 8 deep brain structures: amygdala, caudate head, putamen, globus pallidus, thalamus, hippocampus, midbrain, and pons.

3. SPATIAL STRUCTURE

The standard analysis of brain images, which uses statistics calculated voxel-by-voxel across individuals in each group, requires multiple-comparison adjustments to account for the calculation of up to hundreds of thousands of test statistics, one for each voxel location. Moreover, the individual test statistics do not include any information from neighboring voxels that might be experiencing similar neuronal activity. Spatial statistical modeling (e.g., Cressie 1991), referred to as kriging in the geostatistical literature, can overcome both of these deficiencies when interest is focused on specific brain structures.

3.1 Voxel Correlations

Coordinates for brain maps are often defined using the following neurological convention for the three axes. When the head is viewed from behind, the x direction is left to right, the y direction is back to front, and the z direction is bottom to top. A *transaxial* slice of an image consists of all (x,y) voxel locations for a fixed vertical location, z. A *coronal* image consists of all (x,z) voxel locations for a fixed back-to-front location, y. A *sagittal* image consists of all (y,z) voxel locations for a fixed horizontal location, x. Figure 2 contains typical transaxial SPECT images. The slices contain baseline intensity counts on the left and treatment – baseline differences in counts on the right. The figure illustrates that small neighborhoods tend to have very similar intensities and differences in intensities and that voxels more distant from one another show less similarity.

[Insert Figure 2]

The tendency for strong neighborhood correlations is further demonstrated in Figure 3. This figure shows scatterplots of pairs of residuals from a three-dimensional quadratic surface fit to treatment – baseline differences from the left thalamus of one control subject. Pairs of residuals d mm apart are graphed in each component of the figure, with one of the residuals in a pair arbitrarily assigned to the ordinate and the other to the abscissa. Differences are used in this analysis because of the interest in determining changes in brain activity due to the effects of treatment with physostigmine. Differencing also removes substantial within-subject variability.

The use of residuals eliminates gross trend surfaces that could account for large positive spatial correlations. Kitanidis (1993) recommends detrending so that the *generalized* covariance function, which is the critical component of the spatial correlation function, can be more readily identified. It is clear from the figure that, even after differencing and trend removal, residuals from voxels in close proximity to one another are strongly correlated. The magnitude of the correlation r decreases as the distance between the pairs of voxel locations increases from 2 mm to 7 mm.

[Insert Figure 3]

The strong correlations among differences in intensity counts for voxels in close proximity that are illustrated in Figure 3, even with spatial gradient removal, are typical of all the structures examined in this study. The greatest distances in these structures at which correlations tend to become negligible are approximately 5 – 7 voxels, 10 – 14 mm. The changes in spatial correlations with voxel distances are quantified in the next section.

3.2 ROI Semivariogram Model Fits

Sample semivariogram values were calculated in half-voxel bins from residuals of quadratic fits to the treatment – baseline differences in normalized count intensities, separately for each of the 16 ROIs. The sample semivariogram values for the subjects in each group were then averaged because the primary interest in this study is group comparisons. To ensure that estimated semivariogram matrices were conditionally negative definite, theoretical semivariogram models were fit to the sample semivariograms. A number of potential semivariogram models were discarded because they produced negative estimated nugget parameters and consequent numerical instability in computing algorithms. An example is the spherical semivariogram model, which contributes to these problems because it is linear near the origin. The sample semivariogram values for each of the groups in this study were well fit by the Gaussian semivariogram model

$$\gamma(d) = \begin{cases} 0, & d = 0 \\ \theta_1 + \theta_2 \left\{1 - \exp\left(-d^2/\theta_3^2\right)\right\}, & d > 0. \end{cases} \tag{1}$$

Figure 4 contains sample semivariogram plots for the left caudate of the Syndrome 3 subjects. The jumpy behavior of individual subject semivariograms is evident, as is the much smoother behavior of the average semivariogram. Also shown are the sample semivariogram averages and the Gaussian model fit. Each of the 16 ROIs examined in each of the syndrome and control groups produced semivariogram plots similar to Figure 4.

[Insert Figure 4]

The *nugget* parameter θ_1 in equation (1) provides a very interesting interpretation in the context of brain imaging. The nugget typically quantifies variability associated with microscale variation and measurement error. Microscale variation refers to uncontrolled variation that affects pairs of measurements at distances smaller than the smallest distance in a database, 1 voxel (2 mm) in the Gulf War SPECT study. In geostatistical applications nuggets tend to be large and meaningful, implying a lack of smoothness in soil measurements. Consequently, the Gaussian semivariogram model is infrequently used in geostatistical applications because its use implies smoothly varying measurements. For these SPECT data, however, not only does the Gaussian model provide good fits to the sample semivariogram values but also estimated nuggets are quite small. All are nonnegative but none are significantly different from 0 using z scores calculated from the estimated nugget and its estimated standard error. Because of the differentiability of the Gaussian semivariogram function, this implies that there is strong continuity in nearby voxel intensities for the subjects in the Gulf War SPECT study.

The size of the nugget variation can be better appreciated by comparing the nugget to the *sill*, $\theta_1 + \theta_2$ in equation (1). The sill represents the variability of uncorrelated treatment - baseline differences and is indicated in the figure by the plateau at large voxel distances. The sill is an asymptotic limit for the Gaussian semivriogram models but 95% of its value is attained at a distance of $\sqrt{3}\theta_3$, where θ_3 is referred to as the *range* parameter. The estimated nugget, 3.3, is only 2.2% of the estimated sill 149.2.

The third parameter in equation (1) is the *range* parameter θ_3. For the Gaussian model, it is often concluded that pairs of measurements are essentially uncorrelated at distances greater than the practical range, $\sqrt{3}\theta_3$. Depending on the semivariogram model fit, the range is also the

approximate distance at which the fitted semivariogram values attain or are within 5% of the sill. For this ROI, the practical range was calculated to be $\sqrt{3}\times 3.29 = 5.78$, or approximately 6 voxels. This is one of the largest estimated range parameters for all the ROIs examined in this work. Hence, count normalized intensities for the Gulf War SPECT data can be treated as statistically uncorrelated if the distances between their locations exceed 6 voxels, 12 mm. This finding leads to an improved approach for analyzing group differences based on average count intensities in blocks of contiguous voxels within each ROI.

4. BLOCK SELECTION AND PREDICTION OF BLOCK AVERAGES

The conclusion that neighboring voxels are experiencing similar activity whereas voxels approximately 12 mm apart are uncorrelated suggests group comparisons based on averages of intensity counts from blocks of neighboring voxels. Averages of contiguous voxels within blocks should produce better estimates of the mean activity in the block than do individual voxels. In addition, with the calculation of appropriate prediction variances, blocks within a ROI can be appropriately weighted to produce an average activity for the ROI, separately for each subject. While there are many issues related to selecting and weighting block averages within a ROI that this study cannot comprehensively address, the following discussion details one procedure for obtaining appropriate blocks of contiguous voxels that was effective in making comparisons among the 4 groups of subjects in this study.

For the subcortical gray-matter structures under investigation in the Gulf War SPECT study, blocks were defined with the following constraints: blocks in a ROI contain mutually exclusive voxel locations, blocks were identically defined for all 39 subjects, and the geometric centers of the blocks were at least as far apart as the maximum practical range for the group semivariogram models fit to the ROIs, 6 voxels.

For each ROI, blocks were created by examining the three-dimensional geometry of the structure and selecting contiguous voxel locations that were as convex as each structure permitted. Approximate convexity was achieved by visually examining the voxel centers in each block to ensure, to the greatest extent possible, that pairs of voxel centers could be connected with straight lines that lie wholly within the block. This choice of block geometry facilitated the use of Euclidian distance

within both the blocks and the ROIs, which were approximately convex for most of the structures examined in this study. The number of blocks created per structure ranged from 2 in each side of the amygdala and hippocampus, small deep-brain structures, to 6 in each side of the pons, putamen, and thalamus, which are larger deep-brain structures. That this method of defining blocks does indeed produce approximately uncorrelated block averages was examined by calculating semivariogram values for the block averages. Figure 5 shows the result of those calculations. Due to the small number of blocks in these deep-brain structures, block semivariogram values are shown for a larger region in the left middle frontal gyrus. Between 70 and 100 pairs of blocks are used in calculating the semivariogram values shown in Figure 5. The average semivariogram values for the 4 groups do not show any consistent pattern and the averages are relatively flat, as would be expected for semivariogram values from uncorrelated data. This is the same pattern found in the deep-brain structures that have many fewer blocks from which to calculate the semivariogram values.

[Insert Figure 5]

The calculation of block averages is based on the following model for normalized treatment - baseline SPECT intensity differences within each ROI:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + q(\mathbf{s}) + e(\mathbf{s}), \tag{2}$$

where $Z(\mathbf{s})$ is the treatment – baseline intensity difference at voxel location $\mathbf{s}' = (x, y, z)$; x, y, and z are integer locations using the anterior commissure as the origin; $\mu(\mathbf{s})$ is a nonstochastic, quadratic function of location; $q(\mathbf{s})$ is an L_2–continuous, zero-mean, spatially correlated error process representing inherent SPECT voxel-to-voxel variation; and $e(\mathbf{s})$ is a zero-mean spatial process representing microscale variability and measurement errors. Because the nuggets in the fitted Gaussian semivariogram models are close to zero and not statistically significant, the error term $e(\mathbf{s})$ is deleted from subsequent use of model (2) in this work.

Let $\mathbf{z}(\mathbf{B})$ denote a vector of normalized count intensity differences for an arbitrary set of n locations $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ in a block $\mathbf{B}$. Then a block average can be calculated as $\text{avg}(\mathbf{B}) = \mathbf{w}'\mathbf{z}$,

where $\mathbf{w}$ is a vector of weights determined as follows. Define $\mathbf{B}$ to be the n locations within the block that are shifted half a voxel unit in each direction from the n measured voxel locations (x_i, y_i, z_i) in the block. For the ROIs that are right-half structures, the shifted locations are $(x_i + 0.5, y_i + 0.5, z_i + 0.5)$. The left half-shifted locations are at $(x_i - 0.5, y_i + 0.5, z_i + 0.5)$. The locations in $\mathbf{B}$ are shifted by half a voxel because the predictor using model (2) and a fitted Gaussian semivariogram model that has a zero nugget is an exact interpolator; i.e., it reproduces the actual $z(\mathbf{s}_i)$ at the n measured voxel locations in the block. Shifting the prediction locations by half a voxel allows predictions and their prediction variances to be calculated for a grid of locations within the block, predictions that will be used to form the block and structure averages. These averages over gridded locations approximate the corresponding integral-based averages for the blocks and ROIs.

Denote by $\hat{\boldsymbol{\gamma}}(\mathbf{B})$ the n-dimensional vector whose i^{th} element is the average (across subjects in a group) semivariogram value between the measured location $\mathbf{s}_i$ in the block and the n shifted locations in $\mathbf{B}$. This vector of average semivariogram values is calculated using the fitted Gaussian semivariogram model for the ROI containing the block. Denote by $\hat{\boldsymbol{\Gamma}}$ the $n \times n$ matrix of semivariogram values between pairs of measured locations in the block, again calculated from the fitted semivariogram model for the ROI. Then the universal block-kriging weights (Cressie 1991, Section 3.4) for calculating the block averages are

$$\mathbf{w} = \hat{\boldsymbol{\Gamma}}^{-1}\left[\hat{\boldsymbol{\gamma}}(\mathbf{B}) + \mathbf{X}\left\{\mathbf{X}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{X}\right\}^{-1}\left\{\mathbf{x} - \mathbf{X}'\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\gamma}}(\mathbf{B})\right\}\right], \tag{3}$$

where $\mathbf{x}$ is a 10×1 vector consisting of the average quadratic components of the locations in $\mathbf{B}$ and $\mathbf{X}$ is an $n \times 10$ matrix of the quadratic components of the measured locations in the block. The estimated prediction variance for $\text{avg}(\mathbf{B})$ is

$$\hat{\sigma}_p^2 = 2\mathbf{w}'\hat{\boldsymbol{\gamma}}(\mathbf{B}) - \mathbf{w}'\hat{\boldsymbol{\Gamma}}\mathbf{w} - \hat{\gamma}(\mathbf{B}, \mathbf{B}), \tag{4}$$

where $\hat{\gamma}(\mathbf{B}, \mathbf{B})$ is the average of the n^2 semivariogram values $\hat{\gamma}(\mathbf{s}_i, \mathbf{s}_j)$ for the n locations in $\mathbf{B}$.

Spatial block averages were found to be very similar when neighborhood sizes were varied,

including when the entire structure was used as one neighborhood for each of the blocks in the ROI. Overall conclusions from several investigations of alternative neighborhood sizes were that the method chosen in Section 4.1 assures zero or negligible correlations among predicted block averages within an ROI, provides essentially the same predicted block average as predictions from differing neighborhood sizes, produces approximately the same prediction variance in numerically stable kriging equations using (3), and improves the stability of the kriging system matrix when the system of equations in (3) is unstable (e.g., when the nugget parameter estimate is close to zero or negative and the range parameter estimate is large relative to the ROI volume).

5. SYNDROME AND CONTROL GROUP COMPARISONS

During the course of our extensive analyses of the Gulf War SPECT data, it became clear that one of the subjects in the Syndrome 2 group disproportionately influenced the results. Subsequent to this discovery, medical records were examined and it was determined that this subject received chemotherapy for malignant lymphosarcoma prior to the SPECT scanning sessions and that the chemotherapeutic agents he received can cause neurotoxic injury. Since there was a clinical basis for this subject's abnormal intensity counts, he was eliminated from the final analyses.

5.1 Analysis of Region of Interest Averages

The inverses of the block prediction variances calculated from equation (4) were used as weights to form a ROI weighted average ($\text{avg}\,(\mathbf{ROI})_{ij}$) for subject j in group i. Prediction variances (σ^2_{ij}) for the ROI subject averages were also calculated. The ROI subject averages and their prediction variances were then input into a linear model $\text{avg}\,(\mathbf{ROI})_{ij} = \mu + \theta_i + e_{ij}$, where θ_i represents the group mean and e_{ij} is the heteroscedastic within-subject variability for the ROI average, with variance estimated by σ^2_{ij}. A similar model, without the averaging across blocks, is used in the block analysis reported in the next section.

Table 1 provides the ROI group comparisons. Included in the table are the ROIs where significant ($p < 0.05$) average treatment – baseline differences among the groups were found. In Table 1(a), the ROI group comparisons, the left amygdala, the left caudate, and the left thalamus produced statistically significant 4-group F tests for the entire ROI.

The entries in Table 1(a) below the 4-group F tests are pairwise comparisons of group averages using Tukey's studentized range statistic. These comparisons are two-sided with an inequality A < B in the first column of the table indicating that the significant result was due to group A having a significantly smaller average treatment – baseline difference than group B. Recall that normal brain response to physostigmine was expected to produce relatively large, negative treatment - baseline differences. The significant Control < Syn 2 result for the left caudate has an average treatment – baseline difference for the control group that is negative, as expected. The other two significant results have Syndrome 1 and Syndrome 3 average differences that are also negative. All the Syndrome 2 average differences are positive. The Syn 1 < Control comparison has a very large negative average (-12.0) for the Syndrome 1 group and a weak positive average (1.0) for the control group. The positive average is not significantly different from 0 ($p > 0.05$) and the comparison with the average for the Syndrome 1 group is only borderline significant at the 0.05 level ($p = 0.042$).

The positive average treatment – baseline differences for the Syndrome 2 group caused a further investigation of the treatment and baseline intensities. Figure 6 shows boxplots of the subject average intensities for the baseline and the treatment sessions for the left caudate. A major finding of this work is that the Syndrome 2 subjects typically had low average intensities at the baseline session and the physostigmine challenge in the treatment session raised the intensity averages for the subjects in this group. Physostigmine typically had the opposite effect on the other groups. This pattern of change in the four groups was found to occur throughout the brain, in both the deep-brain regions of interest and throughout the cortical regions of the brain. Haley et al. (1997 a-c) previously concluded that the Syndrome 2 group appeared to be the most seriously affected by service in the Gulf War theater of operations.

[Insert Figure 6]

5.2 Analysis of Block Averages

An important concern in this investigation was whether the syndrome effects were large enough to affect an entire ROI or sufficiently localized so that only portions of deep-brain ROIs are affected. This concern was addressed through the analysis of predicted averages of individual blocks within a structure. Even though structure activation could occur across blocks or wholly within a portion

of a block, the analysis of the individual block averages provided important insight into the extent of voxel activity for this SPECT study.

Table 1(b) shows the results for individual blocks within each of the ROIs that are listed at the top of the table. While Table 1(a) indicates that there are significant 4-group average treatment – baseline differences only for the left amygdala, left caudate, and left thalamus, Table 1(b) reveals that there are significant 4-group differences for individual blocks in every one of the ROIs listed. The results again primarily involve Syndrome 2 average treatment – baseline differences being significantly greater than those of other groups. Medical researchers are keenly interested in the visualization of brain imaging analysis using figures similar to Figure 1. The results of Table 1(b) are graphically displayed in Figure 7. It is interesting and important to note that in structures where two blocks are significant, the blocks are contiguous. Hence, the structure effect is due to either the entire structure (e.g., the amygdala, where there are only 2 blocks) or a regional portion of the structure where voxels in neighboring blocks are affected.

[Insert Figure 7]

It is important to note the relationships between the results reported in Tables 1(a) and 1(b). The statistically significant 4-group ROI comparisons in the left amygdala, left caudate, and left thalamus reported in Table 1(a) result from the significant comparisons shown in Table 1(b) for 2 blocks in each of these small structures. The lack of significant 4-group or pairwise differences for the other structures in Table 1(a) is due to the lack of consistent, significant results in two or more blocks in these structures.

5.3 Comparisons with SPM Methods

SPM group comparisons were applied, as is customary, to the whole brain. The ability to examine all the voxel results in the brain, as well as those for contiguous clusters of voxels, is one of the important strengths of this approach. However, there are aspects of an experiment that can render SPM methods less sensitive to group differences than alternative approaches. With the highly variable Gulf War SPECT data, the small group sizes, and the suspected impairment of only small

areas of the deep brain, routine application of SPM proved less effective than the spatially based approaches described above.

Using whole-brain SPM modeling and analysis (SPM2, Wellcome Department of Imaging Neuroscience, University College, London), there were no statistically significant clusters of voxels using 2-group SPM t-test comparisons or 4-group F tests. It might be that the combination of kernel smoothing between high-signal ROIs and neighboring low-signal ROIs, count normalization using whole-brain means containing signal changes, and inclusion of the entire brain with highly varying signal intensities lessens the ability to detect treatment – baseline changes in specific small deep-brain ROIs in the Gulf War SPECT data. No improvement resulted when whole-brain analyses were conducted without kernel smoothing and with median white matter count normalization. It is possible that the variability present in whole-brain SPECT measurements with small group sizes is too great for these methods to overcome. Possibly because of these difficulties, it is becoming increasingly accepted in brain imaging literature to publish SPM results with p-values uncorrected for multiple comparisons when the multiple-comparison-corrected p-values are not statistically significant.

SPM analyses were redone using the registration and count normalization methods proposed in this article on specific ROIs in order to make as direct a comparison as possible between voxel-by-voxel analyses and the spatial modeling approach. Kernel smoothing of signal intensities was not performed and image intensities were normalized using median white-matter counts from the centrum semiovale. Cluster-level, small-volume-corrected p-values were calculated using expected Euler-characteristic calculations derived from Gaussian random field theory (e.g., Worsley et al. 1996). Table 1(c) indicates that only one cluster of voxels in block 2 of the left caudate produced a significant result, a Control $<$ Syn 2 pairwise comparison. As mentioned above, this significant comparison did not occur with kernel smoothing.

6. DISCUSSION

In the analysis of brain imaging data, spatial modeling and analysis offers important advantages over existing voxel-by-voxel approaches. The advantages accrue primarily from the similarity of neuronal activity in contiguous sections of brain tissue, documented in Section 3.2 by the fitted

Gaussian semivariogram models with nuggets very close to 0 and ranges in excess of 6 voxels. This correlation of activation in neighboring voxels occurs because brain tissue tends to activate in regional patterns of various sizes rather than in isolated point activations and the resolution of SPECT scanners is less than the voxel size. Existing approaches that make adjustments based on many tests at individual voxel locations or that use corrected cluster-level p-values without explicitly incorporating the spatial correlation characteristics sacrifice the ability to detect small changes in the presence of high variability and small group sizes – precisely the situation with the Gulf War SPECT data. Explicitly incorporating the intervoxel correlations greatly increases sensitivity to differences in group averages for specifically chosen regions of interest.

In the left amygdala, the left caudate, and the left thalamus, the spatial analyses are able to combine the averages from 2 significant block comparisons to provide a statistically significant 4-group comparison for the ROI, shown in Table 1(a). The SPM analysis in Table 1(c) relies on clusters of relatively small numbers of voxels in each structure, 79 in the caudate. This cluster size is smaller than those of the significant spatial blocks, approximately 115 voxels in each block in the amygdala, 190 voxels in the caudate, and 100 in the thalamus. Also, in the left caudate, a moderate-size structure consisting of 3 blocks, the SPM analysis identifies only a single significant 2-group comparison, corresponding to our block 2. In contrast, the spatial modeling analysis not only finds the same comparison significant using a larger block size, it also finds a second pairwise comparison significant in a different, contiguous block. This in turn results in the significant pairwise and 4-group comparisons for the entire ROI, shown in Table 1(a). Recall too that the spatial modeling block sizes were determined by distances at which block averages can be treated as uncorrelated – a feature that is not included in SPM cluster-level calculations.

In spatial modeling of brain images, the estimated range parameter characterizes an important feature of voxel correlations. By characterizing the distances over which brain activity is highly correlated, analyses can be customized to particular imaging datasets through the selection of block sizes. The method presented in Section 4.1 for selecting blocks within a ROI is a first approach, one that is reasonable and that does not depend on any knowledge other than the geometry of the structures being investigated. Work is progressing, however, on alternatives that might profitably use voxel activity levels to define contiguous clusters of correlated voxels as blocks of an ROI. Empirically creating blocks in this alternative way could better enable medical researchers to assess

the extent to which portions of structures are activated or deactivated. Such alternative blocking strategies might also help clarify how the activation of one portion of the deep brain is connected to activation or deactivation in other portions by creating blocks that have more homogeneous responses to pharmacologic (e.g., physostigmine) challenges.

Enhancements to the spatial modeling approach used in this article have the potential to further improve the analysis of brain imaging data. ROI spatial trend modeling or spatial filtering might offer improved sensitivity to group differences over the quadratic fits used in this article. Another advance would be the application of nonconvex distance metrics for brain structures that do not have simple geometries. Computationally efficient parametric and nonparametric semivariogram model fitting methods could reduce numerical instability when nugget estimates are close to zero and range parameter estimates are large. All of these potential improvements need to be considered with an added time dimension for alternative medical imaging modalities such as functional MRI.

References

[1] Carmack, P.S., Spence, J., Gunst, R.F., Schucany, W.R., Woodward, W.A., and Haley, R.W. (2004). "Improved Agreement Between Talairach and MNI Coordinate Spaces in Deep Brain Regions," *NeuroImage*, 22, 367-371.

[2] Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: John Wiley and Sons, Inc.

[3] Friston, K.J. (2004). "Experimental Design and Statistical Parametric Mapping," in *Human Brain Function, Second Edition*, ed. R.S.J. Frackowiak, K.J. Friston, C.D. Frith, R.J. Dolan, C.J. Price, S. Zeki, J. Ashburner, and W. Penny, San Diego, CA: Academic Press, pp. 599-632.

[4] Haley, R.W., Kurt, T.L., and Hom, J. (1997a). "Is There a Gulf War Syndrome? Searching for Syndromes by Factor Analysis of Symptoms." *Journal of the American Medical Association*, 277, 215-222.

[5] Haley, R.W., Hom, J., Roland, P.S., Bryan, W.W., Van Ness, P.C., Bonte, F.J., Devous Sr., M.D., Mathews, D., Fleckenstein, J.L., Wians Jr., F.H., Wolfe, G.I., and Kurt, T.L. (1997b). "Evaluation of Neurological Function in Gulf War Veterans: A Blinded Case-control Study," *Journal of the American Medical Association*, 277, 223-230.

[6] Haley, R.W. and Kurt, T.L. (1997c). "Self-reported Exposure to Neurotoxic Chemical Combinations in the Gulf War: A Cross-sectional Epidemiologic Study," *Journal of the American Medical Association*, 277, 231-237.

[7] Haley, R.W., Billecke, S., and La Du, B.N. (1999). "Association of Low PON1 Type Q (Type A) Arylesterase Activity with Neurologic Symptom Complexes in Gulf War Veterans," *Toxicology and Applied Pharmacology*, 157, 227-233.

[8] Haley, R.W., Marshall, W.W., McDonald, G.C., Daugherty, M.A., Petty, F., and Fleckenstein, J.L. (2000). "Brain Abnormalities in Gulf War Syndrome: Evaluation with ^{1}H MR Spectroscopy," *Neuroradiology*, 215, 807-817.

[9] Haley, R.W., Maddrey, A.M., and Gershenfeld, H.K (2002). "Severely Reduced Functioning Status in Veterans Fitting a Case Definition of Gulf War Syndrome," *American Journal of Public Health*, 92, 46-47.

[10] Hammers, A., Koepp, M., Free, S., Brett, M., Richardson, M., Labbé, C., Cunningham, V., Brooks, D., and Duncan, J. (2002). "Implementation and Application of a Brain Template for Multiple Volumes of Interest," *Human Brain Mapping*, 15, 165-174.

[11] Hom, J., Haley, R.W., and Kurt, T.L. (1997). "Neuropsychological Correlates of Gulf War Syndrome," *Archives of Clinical Neuropsychology*, 12, 531-544.

[12] Kang, H.K., Mahan, C.M., Lee, K.Y., Murphy, F.M., Simmens, S.J., Young, H.A., and Levine, P.H. 2002. "Evidence for a Deployment-Related Gulf-War Syndrome by Factor Analysis," *Archives of Environment Health*, 57, 61-68.

[13] Kitanidis, P.K. 1993. "Generalized Covariance Functions in Estimation," *Mathematical Geology*, 25, 525-540.

[14] Lancaster, J., Woldorff, M., Parsons, L., Liotti, M., Freitas, C., Rainey, L., Kochunov, P., Nickerson, D., Mikiten, S., and Fox, P. 2000. "Automated Talairach Atlas Labels for Functional Brain Mapping," *Human Brain Mapping*, 10, 120-131.

[15] Spence, J., Carmack, P., Woodward, W.A., Gunst, R.F., Schucany, W.R., and Haley, R.W. (2006). "Using a White Matter Reference to Remove the Dependency of Global Signal on Experimental Conditions in SPECT Analyses," *NeuroImage*, to appear.

[16] Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain.* Stuttgart: Georg Thieme.

[17] 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 Signals in Images of Cerebral Activation," *Human Brain Mapping*, 4, 58-73.

Table 1. Tukey Multiple Comparison Significance Probabilities (p-values) for Group Comparisons of Average Treatment - Baseline Differences.

Group Comparison	Left Amygdala	Right Amygdala	Left Caudate	Right Hippo-campus	Right Pons	Right Putamen	Left Thalamus
(a) Specified Regions of Interest, Spatial Modeling Averages							
4-Group F Tests	0.015		0.001				0.002
Control < Syn 2			0.009				
Syn 1 < Syn 2							0.003
Syn 3 < Syn 2	0.013		0.003				
Syn 1 < Control							0.042
(b)Individual Blocks in Specified ROIs, Spatial Modeling Averages (block numbers in parentheses)							
4-Group F Tests	0.016(1)	0.026(2)	0.010(1)	0.014(2)	0.032(1)	0.007(3)	0.002(3)
	0.030(2)		0.002(2)				0.001(5)
Control < Syn 2	0.028(2)		0.031(1)				
			0.012(2)				
Syn 1 < Syn 2				0.029(2)		0.026(3)	0.003(3)
							0.001(5)
Syn 3 < Syn 2	0.010(1)		0.025(1)				
			0.006(2)				
Syn 1 < Control						0.028(3)	0.010(3)
Syn 2 < Syn 3					0.025(1)		
(c) Clusters Obtained from Separate Voxel-by-voxel SPM Analysis of Each ROI. (small-volume-corrected; block number in parentheses)							
4-Group F Tests							
Control < Syn 2			0.036(2)				

Figure Captions

Figure 1. Deep-brain Structures Targeted for Investigation in the Gulf War SPECT Study. This graphic illustrates the 8 deep-brain structures of primary interest on the left side of the brain (red) superimposed in the right hemisphere of the brain (gray). Some of these structures border the edge of the hemisphere and some are interior to it. Structures in the right hemisphere are similarly located.

Figure 2. Transaxial Slice of a Typical Baseline SPECT Brain Image and the Corresponding Treatment – Baseline Difference Image. The color scale shows low intensity counts or differences in red, intermediate in blue, and high in green. Voxels in small neighborhoods tend to have very similar intensities and differences, while voxels more distant from one another tend to show less similarity.

Figure 3. Scatterplots of Treatment – Baseline Quadratic-fit Residuals for Pairs of Voxels in the Left Thalamus of a Control Subject. One of each voxel pair is assigned to the abscissa and one to the ordinate. Distances between pairs of voxel locations increase from 2 mm to 7 mm as the scatterplots are viewed from upper left to lower right. Pearson correlations r are shown above the scatterplots. The vertical and horizontal dashed axes cross at the origin.

Figure 4. Sample Treatment – Baseline Semivariogram Values from the Left Caudate of the Five Syndrome 3 Subjects, with Average Semivariogram Values and a Gaussian Model Fit. Subject semivariogram values are open circles, connected by dotted lines. The group averages are solid circles. The fitted Gaussian semivariogram model is represented by the solid curve.

Figure 5. Block Average Treatment – Baseline Semivariogram Values for Each of the Syndrome and Control Groups from the Left Middle Frontal Gyrus, with Average Semivariogram Values Across the Groups. Group averages are numbered; the averages across groups are indicated by the black squares.

Figure 6. Boxplots of Baseline and Treatment Normalized Intensities in the Left

Caudate.

Figure 7. Statistically Significant Differences in Block Averages for the Pairwise Comparison of Syndrome 2 with Other Syndrome and Control Groups. Each ROI is shown in a different color. The dark shade of each color shows the location of the significant blocks within the ROI. The amygdala only consists of two blocks, both of which are significant, so it does not have two shades.

Figure. 1: Deep-brain Structures Targeted for Investigation in the Gulf War SPECT Study. This graphic illustrates the 8 deep-brain structures of primary interest on the left side of the brain (red) superimposed in the right hemisphere of the brain (gray). Some of these structures border the edge of the hemisphere and some are interior to it. Structures in the right hemisphere are similarly located.

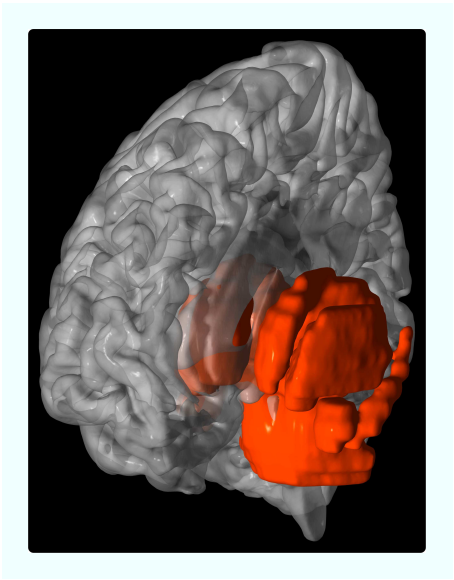

Figure. 2: Transaxial Slice of a Typical Baseline SPECT Brain Image and the Corresponding Treatment – Baseline Difference Image. The color scale shows low intensity counts or differences in red, intermediate in blue, and high in green. Voxels in small neighborhoods tend to have very similar intensities and differences, while voxels more distant from one another tend to show less similarity.

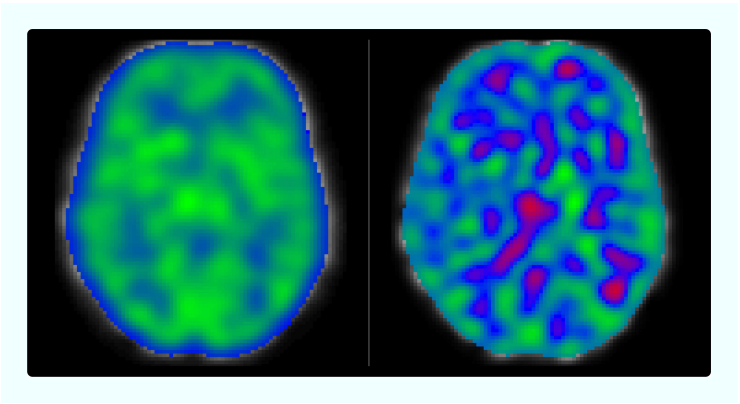

Figure. 3: Scatterplots of Treatment – Baseline Quadratic-fit Residuals for Pairs of Voxels in the Left Thalamus of a Control Subject. One of each voxel pair is assigned to the abscissa and one to the ordinate. Distances between pairs of voxel locations increase from 2 mm to 7 mm as the scatterplots are viewed from upper left to lower right. Pearson correlations r are shown above the scatterplots. The vertical and horizontal dashed axes cross at the origin.

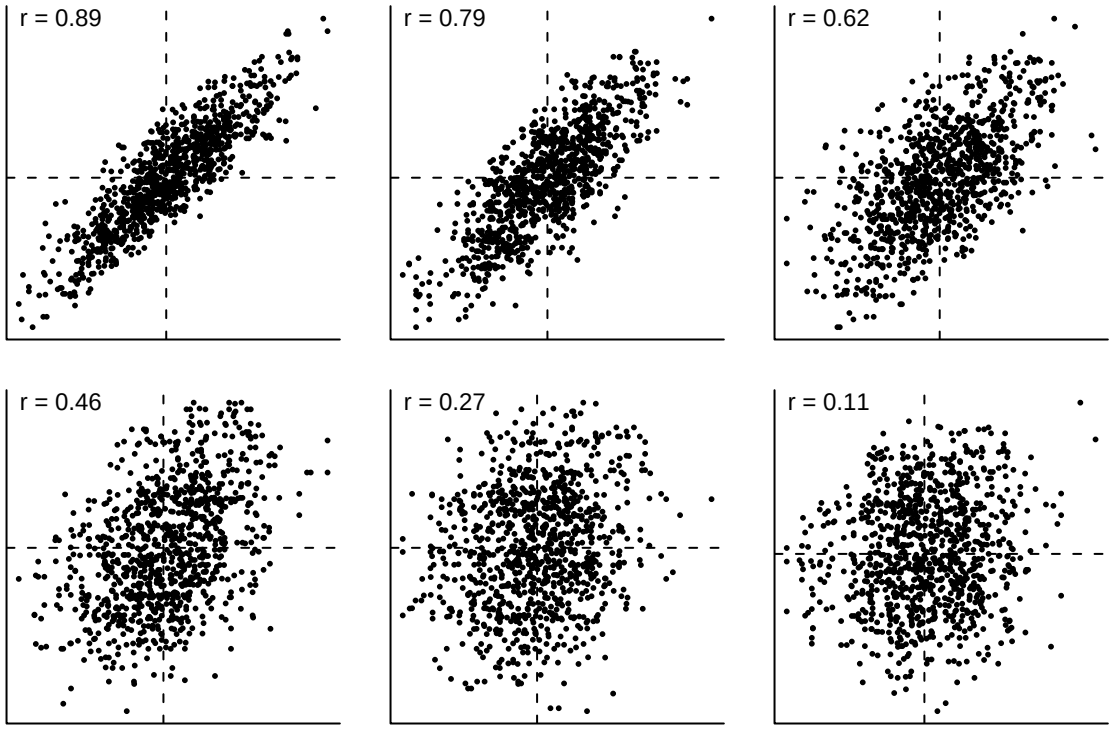

Figure. 4: Sample Treatment – Baseline Semivariogram Values from the Left Caudate of the Five Syndrome 3 Subjects, with Average Semivariogram Values and a Gaussian Model Fit. Subject semivariogram values are open circles, connected by dotted lines. The group averages are solid circles. The fitted Gaussian semivariogram model is represented by the solid curve.

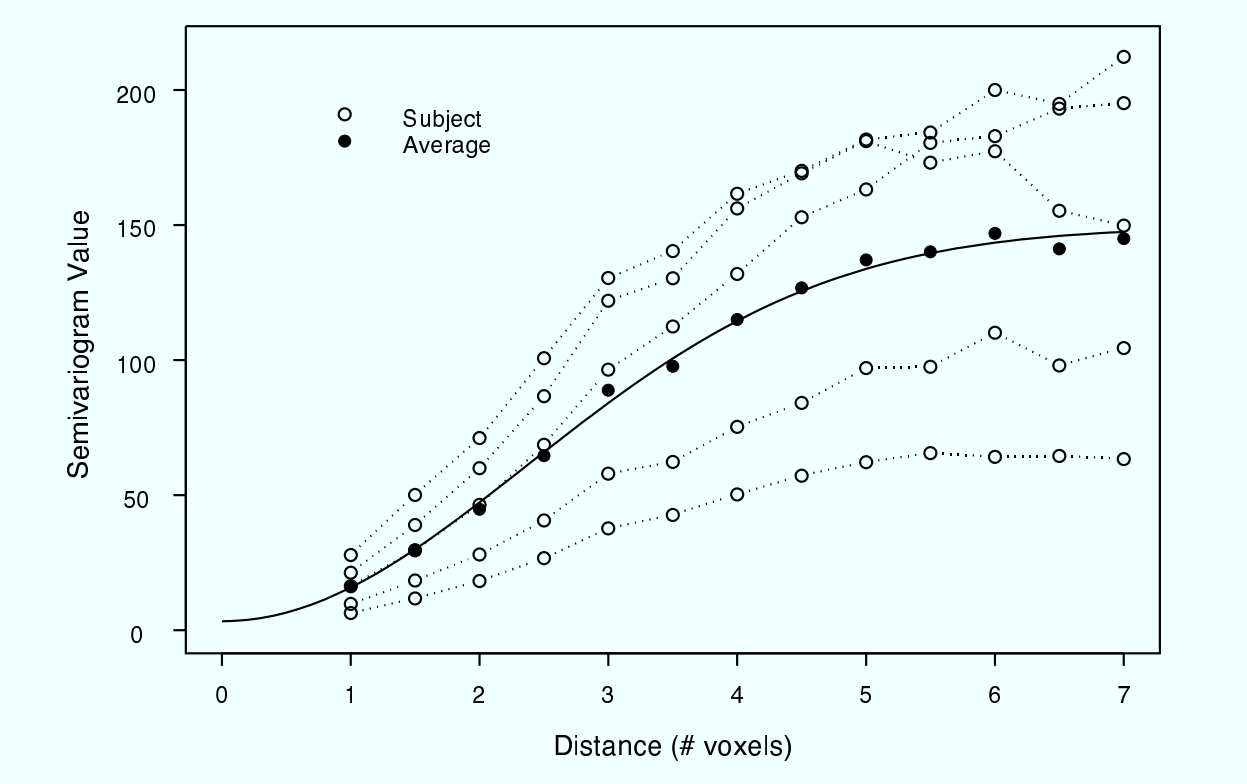

Figure. 5: Block Average Treatment – Baseline Semivariogram Values for Each of the Syndrome and Control Groups from the Left Middle Frontal Gyrus, with Average Semivariogram Values Across the Groups. Group averages are numbered; the averages across groups are indicated by the black squares.

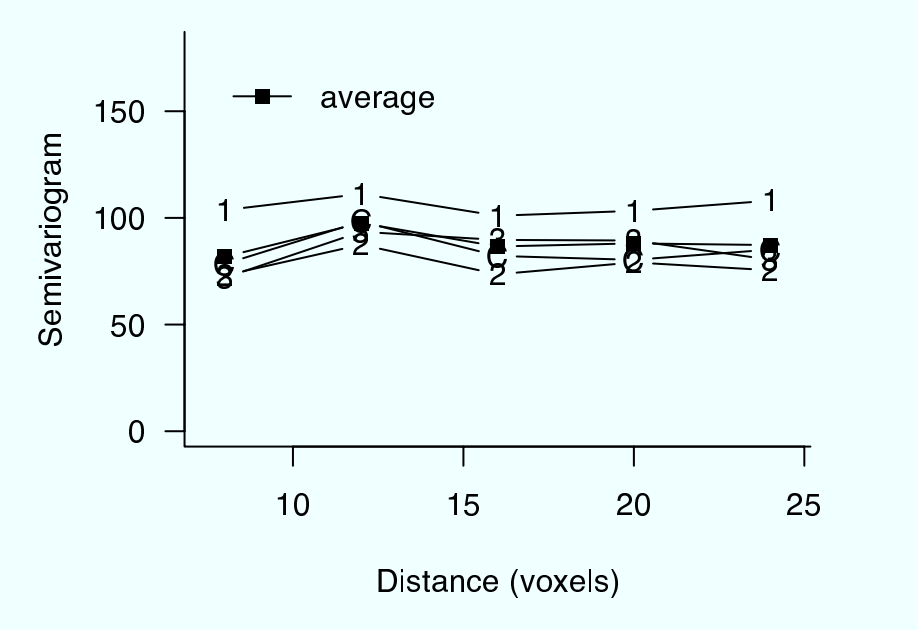

Figure. 6: Boxplots of Baseline and Treatment Normalized Intensities in the Left Caudate.

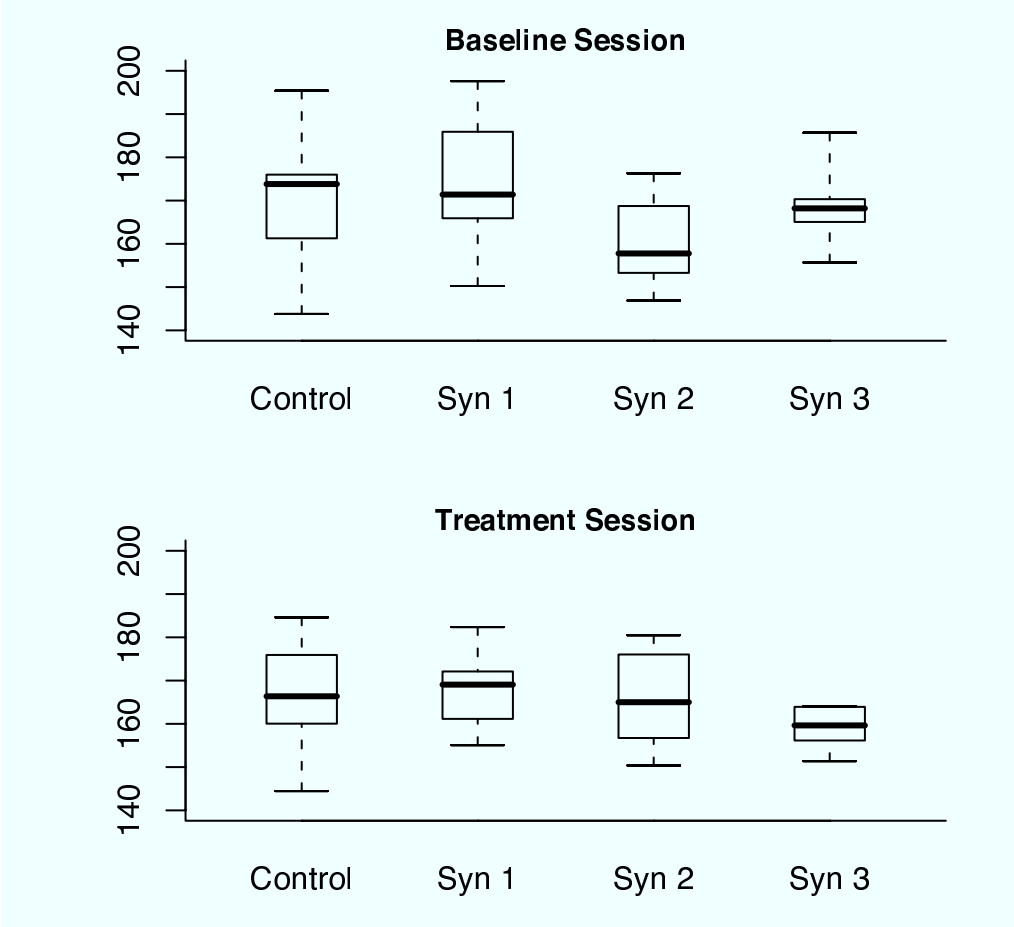

Figure. 7: Statistically Significant Differences in Block Averages for the Pairwise Comparison of Syndrome 2 with Other Syndrome and Control Groups. Each ROI is shown in a different color. The dark shade of each color shows the location of the significant blocks within the ROI. The amygdala only consists of two blocks, both of which are significant, so it does not have two shades.