The Use of Bootstrapping to Measure Image Differences in fMRI Data

by

John Wesley Averick

(Under the Direction of Nicole Lazar)

Abstract

Schizophrenia is a severe mental disorder that affects millions of people and is subject to much research. Functional magnetic resonance imaging (fMRI) is a tool that provides, by observing changes in blood oxygenation, an indirect measure of brain activation. As schizophrenia is known to be partially hereditary, a study was conducted on patients, their relatives, and a control group in order to measure differences in brain activation patterns during the performance of an antisaccade task. In this master's thesis, the processes of bootstrapping and distribution construction are used in an attempt to assess the differences in brain activation among the three groups. The results indicate differences both between the control group and the relatives and between the controls and patients, but no definite evidence that the two differences coincide. Simulated data sets are used to confirm the efficacy of the methodology used. Speculation is given as to incongruities with previous analyses of the data.

KEYWORDS: BOOTSTRAP, FMRI, RESAMPLING, FISHER'S METHOD, SCHIZOPHRENIA

The Use of Bootstrapping to Measure Image Differences in fMRI Data

by

John Wesley Averick

A Thesis Submitted to the Graduate Faculty

of the University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF SCIENCE

Athens, GA

2013

The Use of Bootstrapping to Measure Image Differences in fMRI Data

by

John Wesley Averick

Approved:

Major Professor: Nicole Lazar

Committee: Jaxk Reeves

Lynne Seymour

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

May 2013

Contents

List of Tables

List of Figures

1 Introduction

1.1 What is fMRI?

Magnetic resonance imaging (MRI) is a type of medical imaging that uses strong magnetic fields to generate visualizations of internal body structures. Functional magnetic resonance imaging (fMRI) is a form of MRI that is used specifically for the purpose of viewing the human brain and measuring its activation. The process is done non-intrusively and requires neither surgery nor ingested chemicals. First used by Seiji Ogawa of AT&T Bell Labs in 1990 to measure the flow of blood to the brains of test rats (Ogawa et al., 1990), fMRI is able to roughly display the location and intensity of brain activation due to the effect said activation has on the oxygenation of the blood surrounding it.

The Physiology and Mechanics of fMRI

A neuron is a specialized type of cell found throughout the human body, and it forms the basic building block of the central nervous system. Through its sensitivity to electricity, a neuron is able to effectively transfer signals to other neurons through interconnected chains to specific parts of the body, which then perform the task requested by the signal. Through this system, animals are capable of feeling pain, moving limbs, and having thoughts (Kandel et al., 2000). These, of course, are only

the most elementary functions of the nervous system. The brain, as the primary component of the nervous system, is replete with neurons.

When a neuron fires (is electrically excited), it will remain in that state unless affected by outside stimuli. More energy is required to reset the neuron back to its unexcited state. This energy is provided roughly two seconds later by oxygen ions passing across the neuron's membrane (Huettel et al., 2009). These oxygen ions are carried to the neuron via the circulatory system in the form of hemoglobin, which accompanies influxes of glucose. In order to more efficiently maintain the stability of the nervous system, the blood-flow to the neurons in need of energy is increased locally, rather than system-wide (Huettel et al., 2009). As such, if one particular area needs more glucose while another area does not, both needs can be satisfied. The oxygen ions are used up in the burning of the glucose, but there is often more oxygen than is needed. This excess of charged oxygen ions affects the magnetic properties of the blood in that locality (Huettel et al., 2009).

The magnetic changes caused by the oxygen ions are minuscule, undetectable under the Earth's normal magnetic field, which normally remains beneath sixty microtesla. However, a much more intense magnetic field can throw these minute fluctuations into focus. An MRI machine is a device capable of doing so. The primary component of the machine is essentially a large magnet, which can produce a magnetic field much stronger than what is found in nature, typically 1.5 Tesla or stronger for human subjects (Huettel et al., 2009). When such a strong field is present, the magnetic poles of each individual atom in the human brain are affected.

Normally, the poles of each molecule are oriented in a variety of directions, depending on numerous microscopic forces. In the presence of a strong magnetic field, these poles are forced into an identical orientation, creating a more uniform magnetic signal throughout the brain (Huettel et al, 2009). This uniform magnetic signal makes possible the detection of the otherwise small alterations caused by neuronal activation.

Because the increase in oxygen is responsible for these magnetic fluctuations, the measurements taken by the MRI machine are referred to as the blood-oxygen-level-dependent (BOLD) hemodynamic response (Huettel et al., 2009). These measurements can be taken repeatedly over a short period, providing a working time series of the brain's blood oxygenation. Through this process, it is possible to identify the regions of the brain that are activated in particular tasks.

Unfortunately, measuring the BOLD response at every neuron is impossible. The response itself is not precise enough. The increase in oxygen can only be detected in the general area of the neuron in question. Similarly, there is a limit to the fMRI machine's ability to spatially gauge magnetic resonance throughout the subject. The unit of spatial measurement, when displayed visually, is known as a voxel (volumetric picture element). A voxel is a three-dimensional analogue to a pixel: it is a measure of image resolution. The dimensions of the voxel are arbitrary, and can be set by the operator of the MRI equipment, but there are generally several million neurons within a single voxel (Logothetis, 2008). The size of the voxel can have a large impact on the reliability of the results. The result of a successful fMRI scan is a fully three-dimensional image of the human brain over time, the activation of each voxel represented by a voxel intensity value. The data, once collected, can be arranged into "slices", one-voxel thick groups of data that can provide a two-dimensional image of the brain and its areas of activation at a given depth within the organ.

1.2 What is Schizophrenia?

Schizophrenia is a severe mental disorder that, according to the World Health Organization, affects at least twenty-nine million people around the globe (Barbato, 1999). Despite its severity, it is estimated that at least half of this number are not adequately

treated. Ninety percent of this half can be found in developing nations, where the potential for sufficient care is low and the chance of social ostracism high (Barbato, 1999). In addition to the difficulties brought about by the disorder, sufferers are also much more likely than non-sufferers to experience both clinical depression and substance abuse. Schizophrenia patients have an estimated 25% likelihood of clinical depression, a 25% probability of illegal drug abuse, a 30% rate of alcoholism, and over a 50% probability of nicotine addiction (van Os and Kapur, 2009). The pressure associated with these factors seems to coincide with the one-in-twenty suicide rate among those afflicted (van Os and Kapur, 2009).

The condition has consistently shown itself to be prevalent among disadvantaged social and ethnic groups, however the exact cause of this correlation is a matter of heavy debate (Barbato, 1999). What is more clear, though, is that the domestic environment of the sufferer has a heavy influence on the development and realization of the disorder. Those with a home environment that exhibits traditionally negative characteristics (including but not limited to open hostility) are significantly more likely to develop schizophrenia than those with more benevolent surroundings (Barbato, 1999). Underneath all of these statistics, though, is the fairly recent discovery of a genetic factor in the occurrence of schizophrenia (Kendler et al, 1993).

Symptoms

Schizophrenia is one of the most widely-known mental disorders among laypeople. Unfortunately, genuine understanding of the condition is not as prevalent. In the media and the popular consciousness, schizophrenia is often confounded with dissociative identity disorder (colloquially referred to as "split personality disorder"). This confusion likely resulted from the etymology of the word "schizophrenia" itself, which comes from the Greek roots for "to split" and "mind". In truth, these conditions have little to nothing in common beyond their status as mental ailments.

According to the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (referred to as the *DSM-IV* for abbreviation), the primary professional source for the diagnosing of mental illnesses, schizophrenia is divided into five subtypes. Each subtype has its own distinctive characteristics, representing common groups of symptoms among sufferers. The major signs sought by mental health professionals include delusions (emphatic belief in clearly untrue concepts), hallucinations (seeing or hearing things that are not there), disorganized speech (lack of coherency), and grossly disorganized or catatonic behavior (disruptive or unresponsive states, usually in regard to social interaction) (American Psychiatric Association, 2000).

In addition to these major symptoms, the *DSM-IV* groups three other comparatively minor (and lesser known) indicators under a blanket label of "negative symptoms". The first of these is affective flattening, also known as blunted affect. Those suffering from blunted affect exhibit a lack of emotional reaction, but unlike catatonics do provide basic cognitive responses to stimuli. A second negative symptom is alogia, which manifests as the consistent use of blunt, terse, and undetailed responses. Implied questions will be ignored, so "Are you going somewhere?" will receive a "Yes" or "No" answer despite the universal implication of "If so, where are you going?". Avolition, the lack of motivation or drive to accomplish goals, is the final negative symptom (American Psychiatric Association, 2000).

Diagnosis of schizophrenia is only made if the patient meets certain criteria. Given the severity of the disorder, this list of criteria has been made protracted and exacting in order to prevent false diagnosis. The patient must exhibit at least two of the above symptoms (counting any combination of the negative symptoms as one) for a significant portion of a one-month time period. During this time, a "major area of functioning" (American Psychiatric Association, 2000, 148), meaning occupation, marital relationship, academic performance, or another significant aspect of the patient's life, must be adversely affected. With this established, monitoring of the

subject must confirm that subsequent disturbances continue for at least a six-month period before a definite diagnosis can be made.

Subtypes

The most common subtype of schizophrenia is the paranoid type. This subtype is primarily characterized by delusions and auditory hallucinations. While these delusions are often obsessive, disorganized speech or behavior is absent, as is catatonia. These patients are coherent in thought and expression, but the ideas they put forth can be confusing and disconcerting to others (American Psychiatric Association, 2000). The catatonic type of schizophrenia is marked by the various unresponsive behaviors that fall under the description of catatonia, including but not limited to resistance against mobility, complete laxness in the limbs, or "waxy" behavior in which limbs can be moved but will stay in place once movement ceases, regardless of discomfort (American Psychiatric Association, 2000). The disorganized type coincides with disorganized speech and blunted affect (American Psychiatric Association, 2000).

The preceding three subtypes describe nearly all well-defined cases of schizophrenia. Given the near-infinite complexities of the human brain, however, it is not entirely unusual for a patient to fail to meet the strict criteria listed for each (American Psychiatric Association, 2000). The undifferentiated type exists for those patients who display symptoms across one or more of the other subtypes. For example, a patient may suffer from powerful delusions and exhibit grossly disorganized behavior (American Psychiatric Association, 2000). Rather than diagnose the patient with both the paranoid and disorganized subtypes, the patient is said to suffer from undifferentiated schizophrenia. The final schizophrenia subtype, the residual type, addresses the least severe cases. These patients either exhibit only the less severe negative symptoms (such as flat affect) or, if they suffer from the more prominent symptoms as well, they

do so in a much less prominent fashion. For example, delusions may manifest simply as a harmless cognitive block, such as insisting that grass is blue. An easily-referenced table of the subtypes and their symptoms is provided in Table 1.1.

Table 1.1: The subtypes of schizophrenia and their symptoms

	Delusions	Hallucinations	Disorganized Behavior	Catatonic Behavior
Paranoid	Yes	Yes	No	No
Disorganized	No	No	Yes	No
Catatonic	No	No	Yes	Yes
Undifferentiated	Maybe	Maybe	Maybe	Maybe
Residual	No/Mild	No/Mild	No/Mild	No/Mild

1.3 Mental Illness and fMRI

As the brain is the organ that governs all human thought, the study of mental illness inevitably leads to the brain. Science has long shown that damage to parts of the brain can change a person's behavior (the account of Phineas Gage is often cited among laypeople), among other important conclusions related to the organ's processes. Before fMRI, however, the actual biology of the brain offered few answers to "adversely affected" states of mind. Post-mortem examinations could be performed, but there was simply no way to observe the living brain without intrusive and potentially dangerous surgical procedures.

The advent of fMRI has changed this. By being able to observe the functioning of the brain and notice quantifiable differences in its activation, researchers have the opportunity to better understand the regions of the brain that might be affected by mental illnesses. The applications of this information in the field of mental health are potentially staggering, and could lead to advances in the more effective treatment of patients. However, the data must first be adequately modeled and interpreted.

Statistical Analysis and fMRI

The results of fMRI scans provide some challenges in terms of statistical analysis and interpretation. The size of the data sets produced can be daunting, with each subject producing several million individual measurements. As a result, automated routines are required to do any analysis of the entire sample. Another practical concern is the fact that no two human brains have an identical shape. In order for direct comparisons between subjects to be reliable, the data sets must each be transformed into a common shape and size. There is also the concern of maintaining a desirable family-wise error rate: voxel-by-voxel comparisons result in hundreds of simultaneous hypothesis tests. Traditional approaches to handling the multiple testing such as using Bonferroni correction push the significance threshold for individual voxels down to levels that are difficult to achieve.

Another issue is noise, variation in fMRI images over time that is not caused by experimental manipulation. This is opposed to signal, which is variation as a result of experimental manipulation (Huettel et al., 2009). Under ideal circumstances, the only variations found over the course of an fMRI time series will be those caused by the stimuli administered by the experimenter. These circumstances would allow the experimenter to easily identify the effects of the stimuli with little or no ambiguity. Such a scenario is not realistic, however. Noise of some magnitude is unavoidable, and can come from several sources. The heat generated by the imaging equipment, as well as imbalances or imprecisions in the alignment of certain components, can result in noise. The latter can be largely avoided by thorough equipment checks prior to imaging. The former scales linearly with increases in the magnetic field, while signal itself scales quadratically, meaning a stronger magnetic field will decrease the appreciable effect of temperature (Huettel et al., 2009). The primary source of noise is physiological, resulting from the bodily processes of the subject. These include but are not limited to increases and decreases in pulse rate, as well as head movement.

In the data set described in the next chapter, attempts are made to address these concerns. The original researchers applied motion correction to remove the effects of head movement, used a Gaussian filter to compensate for individual variations in anatomy, and transformed data from all individuals to the standard Tailarach space (Talairach and Tournoux, 1988) to facilitate comparisons. On our part, to remove any potential effects of "air voxels", voxels outside the brain, we applied a simple filter that removed data points with especially low magnetic resonance signals, which we believed to occur outside the brain. We used two-sample t-tests to remove the time series aspect from the data by giving a single representative t-score for each voxel of each subject. We then used Fisher's method (described in Section 2.4) to combine the brain images of multiple subjects. Through these methods, the large data sets become more manageable in size while still retaining statistical information. In order to limit the complications of multiplicity and the uncertainty of what practical location a single voxel may constitute, we used three measures of image differentiation to test eight larger divisions of the brain for significant activation difference, rather than testing thousands of solitary voxels. We used Bonferroni correction to correct for these multiple, but considerably less numerous, simultaneous hypothesis tests.

Instrumental to our analysis is the process of bootstrapping, a form of resampling. By sampling with replacement from a sample itself, bootstrapping allows us to build an empirical distribution for a random variable based on observed values (Efron, 1979). This method is capable of providing a more accurate p-value as it does not assume that a random variable follows a particular distribution, an assumption that may not hold. Through more accurate p-values, we may better assess the statistical significance of an observed value from a sample.

2 Data and Methods

2.1 The Study of Interest

The subject of our analysis is an fMRI study conducted on a sample of 45 individuals. The main concern of the study was the genetic factor of schizophrenia. Both those with schizophrenia and their close relatives have shown similar difficulties in performing volitional saccade tasks, tasks that require subjects to quickly move their eyes in a manner dictated to them. By observing the regions of the brain in which both groups show significantly different activation from control subjects, the study hoped to more narrowly define the areas of the brain involved in performing eye-related tasks that are adversely affected by schizophrenia (Camchong et al., 2008).

The 45 individuals consisted of three distinct testing groups. The first was made up of seventeen patients, all diagnosed with schizophrenia as defined by the *DSM-IV*. Thirteen of the patients were being given antipsychotic medication at the time of their participation in the study. The second testing group included thirteen biological relatives of the patients: one parent, eleven siblings, and one child. The final testing group was a control. Fifteen subjects with no known history of schizophrenia (personal or hereditary) were recruited through a newspaper ad. Because of the observed differences in brain activation related to left-handedness, all subjects were required to be right-handed (Camchong et al., 2008).

The procedure conducted on the subjects consisted of thirteen alternating time blocks, six designated as task, seven designated as rest. During the task blocks, the subject was presented with the image of a centrally positioned cross, followed immediately by a gray dot to either the left or right. This would occur eight times over a period of 25.2 seconds. The subject was instructed to look in the opposite direction of each dot's appearance as opposed to instinctually looking toward it. Such a task is called an antisaccade task (Camchong et al., 2008). Those diagnosed with schizophrenia have been shown to perform poorly on such tasks compared to control subjects, often looking toward the stimulus (a prosaccade) or hesitating before looking away. It is speculated that the patients' condition reduces their ability to suppress activity at saccade-related neurons (Munoz and Everling, 2004). The rest blocks consisted of the subjects focusing on a single black cross bordered by a square in the center of the display for a period of 22.5 seconds. The fMRI scans are time series over the course of 308 seconds, with a scan occurring every 3.8 seconds. The result is a sequence of 81 images. The brain images produced consist of 38 horizontal slices, each four millimeters thick (Camchong et al., 2008).

After the researchers gathered the data, spatial correction was used to compensate for minor head motion. A full-width half-maximum Gaussian filter was applied to account for individual distinctions in anatomy. The researchers achieved this by convolving the data with a Gaussian distribution of a specified width, with "full-width half-maximum" referring to the difference between the two values at which the distribution reaches half its maximum probability, as illustrated in Figure 2.1. The researchers chose a width of four millimeters. On a voxel-by-voxel basis, the percent change in BOLD signal between the rest and task blocks was calculated for each voxel in each image. These percent changes were used to construct a six-factor random effects linear model for each subject, with a baseline factor, a factor for experimental conditions, a factor to compensate for linear drift, and three factors to

compensate further for head movement (Camchong et al., 2008). The researchers then transformed the image data to Talairach space (Talairach and Tournoux, 1988). Each transformed brain image has a resolution of 40 by 48 voxels. The parameter estimates for experimental condition at each voxel from all subjects were then subjected to a one-sample t-test. The resulting t-map allowed the researchers to identify regions of significant BOLD signal change. These regions, once identified, were each analyzed independently of the rest of the brain. More specifically, a sphere (with an eight millimeter radius) was placed centrally around each region of apparent significant change. Within each of these spheres, the researchers calculated mean intensity changes for each individual, and for each region an ANOVA test was conducted to identify the effect of treatment group on the change.

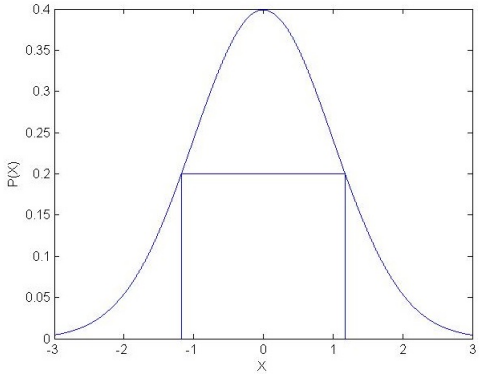

Figure 2.1: An illustration of full width at half maximum for a standard normal distribution. Specifically, the length of the horizontal line segment is the distance between the two values where the distribution function displays half its maximum probability. The vertical lines are a visual aid, helping to more easily compare the full width to a standard deviation.

The results of the study seemed conclusive. Both the patients and their relatives showed decreased brain activation in normally active neural regions, but with less extreme decreases among the relatives. In addition, other regions of decreased activation were present only in the relatives (Camchong et al., 2008). More specifically,

both groups showed decreased activation during task in the cuneus, the insula, the middle occipital and anterior cingulate gyri, and the dorsolateral prefrontal cortex. In these areas, the decrease in activation was less pronounced among the relatives than among the patients. In the lateral frontal and supplementary eye fields, the patients showed decreased activation that was not shared by their relatives (Camchong et al., 2008). However, it may be beneficial to examine the difference between the treatment groups with methods more robust to the assumptions made by an ANOVA test (in particular, the assumption of normal residual distribution). Among these methods we look at resampling techniques, in particular the process of bootstrapping (Efron, 1979).

2.2 Bootstrapping

The problem of generalizing the findings of a sample to the population as a whole is omnipresent in statistics. In the original study, Camchong et al. came to their conclusions based on the ANOVA of results derived from random effects linear models (Camchong et al., 2008). Another approach to the issue is to combine test statistics across multiple patients and test this conglomerate statistic for significance. This method has shown itself to be more sensitive in detecting significant activation than the use of linear models (Lazar et al., 2002). Also, through the use of bootstrapping, we can avoid a potential shortcoming of the ANOVA tests. In the use of ANOVA we assume that errors follow a normal distribution. This assumption does not always hold. Bootstrapping, however, constructs empirical distributions which should more accurately estimate the probability of a particular observance. The outcome of such analysis may prove useful as a companion analysis to linear models.

The bootstrapping procedure involves sampling from the sample itself, with replacement, a number of subjects equal to the size of the original sample (Efron and Tibshirani, 1993). This process can be done with or without regard to treatment groups: each group could sample only from itself, or it could sample from the sample as a whole. This "resample" is then made subject to whatever statistical tests are used on the original sample. The process is repeated a hundred, a thousand, or however many times is desirable. The results from the many resamples will build an empirical distribution for the test statistic, against which the initial result from the original sample may be compared and an accurate p-value obtained (Hesterberg et al., 2005).

While distributions for many measurements are by definition continuous, the distributions constructed by bootstrapping are always discrete (there are, after all, a finite number of ways in which to reassign or duplicate the subjects). Because of this, the more resamples with unique test statistics that are obtained, the more comprehensive the empirical distribution will be and the more reliable the p-values will be, as the distribution will more accurately reflect values a random variable is likely to take. The ideal situation would of course be to bootstrap such as to achieve every possible unique test statistic, as this would create a "complete" (though still discrete) distribution of the variable containing all possible values that the sample can yield. However, even moderate sample sizes can render the number of distinct test statistics so high as to make this computationally impractical (Smyth and Phipson, 2010).

A bootstrap-constructed distribution is unique to the data that it is created from, and shaped entirely by values observable in the sample. This provides certain advantages and disadvantages compared to using more well-known, traditional statistical distributions for analysis. Because of its close relationship with the data, a constructed distribution is more likely to reflect atypical or complicated properties in the population that may not conform to more traditional distribution shapes. This can

result in a more reliable p-value as it will more closely relate a statistic's extremity to other observed values. However, the bootstrap process is limited in that the distribution contains no more information than can be found in the sample itself, and as such may not fully reflect the range of values that a variable is likely to take.

In our analysis, the bootstrapping procedure is applied to the individuals involved in the study. To decrease the effect of random sampling error as a result of the bootstrapping procedure, we use a large number of bootstrap resamples. A number of one thousand is chosen, largely arbitrarily but also based on the necessary computation time involved in the process. In the interest of computational efficiency, rather than bootstrap the original data we are concerned with, an air voxel filter is used to help remove signals that do not result from brain activity and a two-sample t-test is used to summarize the difference between rest signals and task signals at each voxel in each individual. The bootstrapping is then performed with each subject represented by a single brain image of t-scores. The subjects are resampled with replacement and without regard to their original testing group.

This resample is then subjected to a methodology that results in a more compact and easily interpreted representation of the difference in brain images between testing groups. In particular, Fisher's method (described below) is used to combine the fMRI scans of individuals in the same testing group, and three separate image differencing measures are used to gauge the dissimilarities in the combined-brain images from each testing group. The p-values resulting from this process should aid us in identifying areas of the brain where different treatment groups exhibit significantly different BOLD signal activation. The results from the original data, also subjected to this methodology, are then compared to the results generated by the resamples, providing insight into the extremity of the differences recorded.

2.3 The Data: General Observations

Before this procedure is carried out, though, it is important to address issues in the data. The results of previous research allow us to narrow our focus to slices where task-related activation is most likely to be found. The inclusion of slices that have no relevance to the tasks being performed can potentially obscure any results, depending on one's methodology. For the purposes of this discussion, we will focus on five mid-brain cross-sections: the 25^{th} through 29^{th} slices (measured from the bottom slice of the brain). Unfortunately, for one of the schizophrenia patients, there was an obvious error in the scans performed, resulting in posterior portions of the brain being completely masked from measurement across all slices. This renders the patient's data largely unusable, and as such the data have not been utilized in our analysis. Another patient's data were discarded prior to the report's publication "due to insufficient contrast between the pupil and the iris" (Camchong et al., 2008, 1043). Discarding these two, our final sample consists of 43 individuals.

We also wish to address the issue of air voxels, the space outside the brain in the images. Non-physiological noise does cause variation over time in this area, although the magnitude of the variation is smaller as the magnetic resonance signal in air voxels are usually lower than it is in gray matter (Huettel et al., 2009). Even so, it serves no use in the present analysis and, if included, still has the potential to obscure results. As such, we have utilized a simple, conservative filter that removes voxels with especially low magnetic resonance signals, resulting in images less likely to include air voxels.

2.4 The Methodology: Hypothesis Testing

One of the primary interests in the performance of antisaccade tasks is observing how brain activation differs between task and rest states. We felt that the easiest

way to make this direct comparison was through a two-sample Student's t-test, with the signals of each voxel at either state being the two samples. Due to potential artifacts caused by the starting of the fMRI equipment, the first two images are dropped, leaving a 79-image time series for each patient at each slice. Of these, 41 are considered task and the remaining 38 are considered rest, according to the records kept by the researchers. Since it is expected that a voxel at task would have a larger signal than a voxel at rest, the t-test is right-tailed. In the interest of robustness, we choose not to assume that the samples have equal variance. The t-score at each voxel is defined as

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where "$\overline{x}_1$", "n_1" and "s_1" are the sample mean, sample size, and sample standard deviation of the voxel in the task images, respectively, and "$\overline{x}_2$", "n_2" and "s_2" are the sample mean, sample size and sample standard deviation of the voxel in the rest images, respectively. This value is then compared against a Student's t-distribution with the Welch-Satterthwaite equation

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

used to calculate the distribution's degrees of freedom (Welch, 1947). Through these tests we obtain a single p-value for each individual voxel in each patient's brain image. It is notable that this two-sided t-test does not compensate for temporal correlation, treating each voxel in each image of the time series as a wholly independent measurement (Christensen and Yetkin, 2003). In spite of this deficiency, we believe the two-sample t-test will be sufficiently accurate for our purposes.

Our next concern is the amalgamation of the subjects' data into representative images, thus facilitating direct comparison of the treatment groups. Several methods exist to combine t-scores. The most popular, though, is most likely the Fisher method, which possesses "certain statistical optimality properties" in comparison to other methods (Lazar et al., 2002, 550). The Fisher method is based on the following formula:

$$T_F = -2 \sum_{i=1}^{k} \log(P_i)$$

Where "k" represents the total number of subjects in the treatment group. "P_i" represents the p-value from the relevant voxel in the i^{th} subject of the group. "T_F" is the final representative value to occupy a particular voxel for that group. If there is no significant difference in task and rest activation in a voxel for any patient in that group, this statistic should follow a chi-square distribution with 2k degrees of freedom (Fisher, 1950). The result of this amalgamation of data is a single representative brain image for each group for each slice of the brain. Figure 2.2 gives a sample of the results, with control, relative, and patients from left to right. Areas of high but noticeably different activation are visible in all three groups, particularly in the anterior regions of the brain. Whether the differences in these regions are the result of disparities in the brain activation of the treatment groups, or whether they are simply the result of random variation, is the issue which we must address.

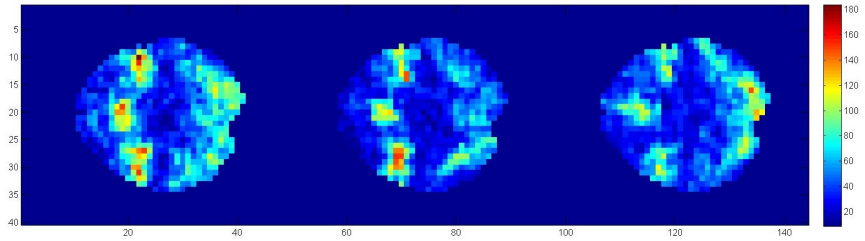

Figure 2.2: Fisher's test results for 28^{th} slice. From left to right: control, relatives, patients.

Thus far, we've concerned ourselves with operations performed at the individual voxel level. However, a purely voxel-by-voxel approach to our analysis is not preferable. The problems we encounter with such an approach are ones of multiplicity and interpretation. Performing hundreds of simultaneous hypothesis tests would (after correcting for multiple comparisons) push the requisite p-value for statistical significance down to a low level, one which few if any of the voxels could ever meet. There are also concerns about how to interpret the results of such a test: we come to the difficult question of what a single voxel represents in practical terms. In reality, it is simply a construct used to compensate for the fact that our technology cannot perfectly replicate the infinite complexity of matter. While a group of significantly active voxels might indicate an area of the brain that is experiencing higher than normal activation, a lone significantly active voxel, while technically a result, is largely meaningless.

For our test statistics, we have chosen three different measures for appraising the distance between two images (Van der Weken et al., 2004, and Wang, 1997). Using multiple measures will give us a comprehensive look at the image differences, and the weaknesses of one method may be compensated for by the others. The three measures we utilize are:

$$d_1(A,B) = \frac{1}{m_1 m_2} \sum_x |A(x) - B(x)|$$

$$d_2(A,B) = [\frac{1}{m_1 m_2} \sum_x |A(x) - B(x)|^2]^{1/2}$$

$$d_3(A,B) = \frac{\sum_x |A(x) - B(x)|}{\sum_x |A(x) + B(x)|}$$

In the above formulas, "A(x)" and "B(x)" represent the paired values between the two groups at each voxel, and "$m_1 m_2$" represents the total number of voxels being compared. These formulas produce a single value for each pair of images.

However, a single global value is unlikely to shed light on our primary concern: in what areas of the brain do the groups differ? To address this, we divide the images into octants (four deep, two across, as illustrated in Figure 2.3), and run the image differentiation measures on each octant separately. This allows us to identify distinct areas of differentiation, as each octant will have its own test statistic for each two-way comparison among the three testing groups. A legitimate criticism one might raise with regard to our eightfold division is that it is no less arbitrary than the division into voxels. While this is true, it is notable that one of our divisions is large enough to generally describe an area of activation in a practical sense, whereas the area described by one voxel is not as easily interpretable.

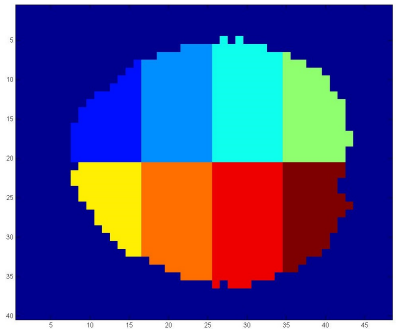

Figure 2.3: An illustration of the eightfold division applied to the brain images. Pictured is the 25^{th} slice, with each differently-colored portion of the brain representing an octant.

With our final test statistics now defined, we may formally state our hypothesis.

H_0: No pairwise comparison of groups reveals significantly different brain activation.

H_A: At least one pairwise comparison reveals significantly different brain activation.

The bootstrapping procedure repeats the process described above one thousand times. Specifically,

1. Resample the subjects' t-maps created by our two-sample t-test. The t-maps are resampled rather than the full original data in the interest of computational efficiency, as a particular subject's t-map will remain unchanged regardless of how the resample arranges or duplicates it.

2. Perform Fisher's test on the resulting resample, combining the subjects assigned to each treatment group into a representative brain image for that group.

3. Use our three image differentiation measures on all possible combinations of the treatment group brain images, calculating a value in each octant separately.

Checks are put in place to ensure that a resulting resample is never duplicated. We resample without regard to groups. For example, a subject from the control group could be randomly resorted into the schizophrenic group or vice-versa and so on. Our null hypothesis is that there is no significant difference in brain activation in any region between the groups so, under that assumption, it would not make a difference which subject appeared in which group, as the result would be roughly the same. However, if the original sample shows differences that are considered extreme under such an assumption, then there is evidence that there are differences among the subjects in each group. From the one thousand resamples, a distribution is constructed for each of the octants for each of the three possible group comparisons in each of the five slices. The proportion of test statistics in a distribution as extreme or more than the one observed in the original data will become the p-value for a given octant.

3 Results

3.1 Concerns

Two examples of the empirical distributions generated by our bootstrapping method are illustrated in Figure 3.1. The values represented are the test statistics produced by the first image differencing measure, specifically from two anterior octants in the pairwise comparison of the controls and the relatives on the 25^{th} slice. The test statistic calculated from the original data is marked with a red vertical line, illustrating its extremity. In these two cases, the brain activation in the octants is found to be significantly different between the two groups. Figures 3.2 and 3.3 show all of the statistically significant differences found between the three groups. The values represented by the colors are the p-values themselves, scaled to highlight octants that show significantly different activation at a family-wise error rate of $\alpha = 0.1$. Bonferroni correction is used to determine the significance threshold. We illustrate higher p-values, indicating regions without a statistically significant difference in activation, with a deep maroon. We illustrate lower p-values, indicating regions that exhibit a statistically significant difference in activation, by a spectrum of colors depending on how close each p-value is to the significance threshold. The p-values close to zero tend toward blue, while the p-values closer to the significance threshold tend toward red. The exact sequence of the color spectrum's relation to the p-values is shown to the right of each image.

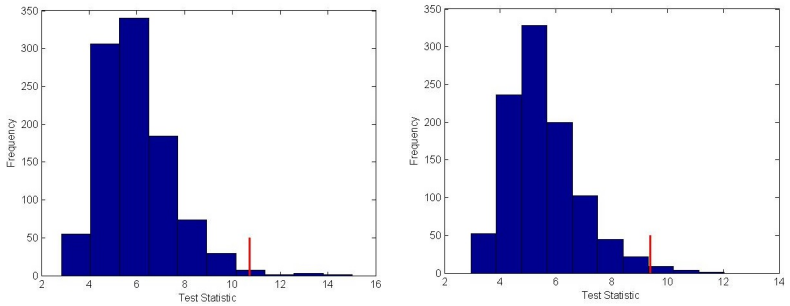

Figure 3.1: Distributions of the test statistics from two brain image octants. Taken from the 25^{th} slice, using the first image differencing measure to compare controls and relatives.

Both the first and third image differencing measures consistently show differences between testing groups. In particular, two anterior octants of the 25^{th} slice were found to have significantly different activation between the control subjects and the relatives of the patients; the third measure found an additional anterior octant in the 26^{th} slice that met this description as well. The same measures found a single anterior octant in the 27^{th} slice that exhibited significantly different activation between the controls and the patients. No significant differences were found in the 28^{th} or 29^{th} slices by any measure, and the second measure did not discover significant differences on any of the five slices.

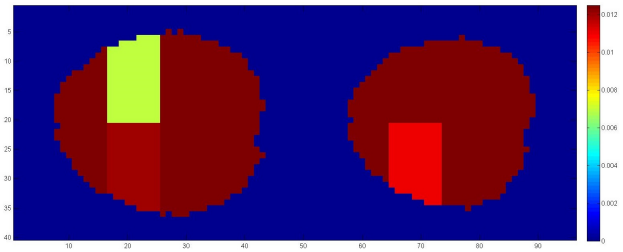

Figure 3.2: Results of the first measure (d_1). From left to right: control/relatives at 25^{th} slice and control/patients at 27^{th} slice.

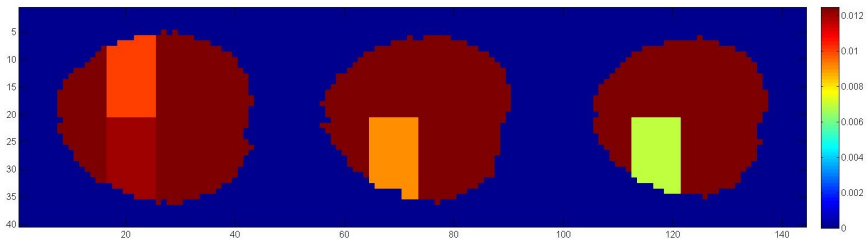

Figure 3.3: Results of the third measure (d_3). From left to right: control/relatives at 25^{th} and 26^{th} slices and control/patients at 27^{th} slice.

These results do not correspond with the findings of Camchong et al. Our methodology has found no regions in which both the patients and their relatives exhibit significant differences from the control subjects. Also, regions of the brain normally extend to multiple slices. While it is possible that the difference between the controls and the patients is confined to the 27^{th} slice, the interpretation of such a small location is not entirely clear. This disparity in results may warrant a closer look at the effectiveness of our methodology.

3.2 Simulation Tests

One way of confirming the suitability of our methodology is to simulate data of our own. To do this, we will fabricate two groups of brain images with similar levels of activation and include regions of higher activation in one. If our methods cannot identify intentionally placed differences in brain activation between the two groups, then our methodology's usefulness is in question. If our methods are successful, then our confidence in them can be substantiated. For this simulation, we return to the t-score maps previously calculated for the experimental subjects. At each slice used in our analysis, we use the resulting t-scores from our two-sample t-tests to construct a distribution of possible values for each voxel. We do this so that the constructed brain images will exhibit t-scores in a range similar to the real data.

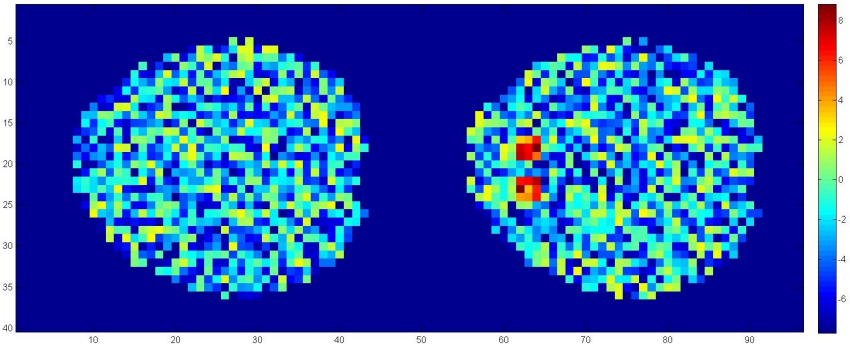

Figure 3.4: An example of two randomly generated t-score brains based on the 25^{th} slice, scaled identically. At right is the brain with two areas of high activation placed in the anterior portion. Two groups of fifteen such brain images will be compared using our methodology to ascertain if it is sensitive enough to detect the differences that we know exist.

We calculated quantiles defining the lower 95% of the distribution, classifying values within this range as non-extreme. Non-extreme values are assigned to groups of voxels in the shape of brain images. These values are uniformly selected from the range designated non-extreme, with the brain shape based on the slices analyzed. The same non-extreme values are then used to construct a second group of brain images, however two anterior octants are each given a three voxel by three voxel cluster of values considered extreme. The extreme values are taken uniformly from the 5% range at the upper extremity of the t-scores' distribution. Figure 3.4 illustrates an example of two constructed t-score brain images. We use the upper extremity alone rather than both the upper and lower extremities because the t-tests used in our methodology are right-tailed. We are interested in voxels that become more active in response to the task, compared to their status at rest. This method of construction creates two simulated samples of fifteen brains, fifteen being an arbitrarily chosen number similar to the size of our sample groups. We observe the results of our methodology upon the constructed data, then construct another simulated dataset, and so on until we have simulated the data thirty times for each slice, thirty being an arbitrarily chosen number, for a total 150 simulations.

With these new data, we may observe whether or not the image differencing measures are capable of finding the difference between the two groups. Figure 3.5 provides an example of the p-values returned in a single analysis of the constructed data. Octants without a significant difference in brain activation are again shown as deep maroon. Octants that do exhibit significantly different brain activation are shown as green. All three measures were capable of identifying the two octants with clusters of extreme values in our simulated data. These same results were reached for all thirty constructions and analyses of brain images for each of the five slices. Every analysis showed consistent detection of the two octants in question and no false positives were ever registered in other octants. Such singular results strongly support our methodology's efficacy.

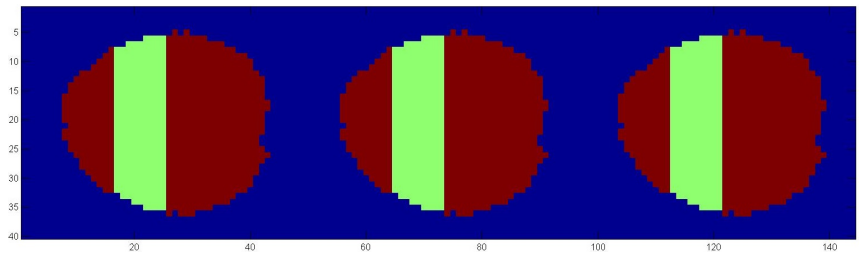

Figure 3.5: Octant test results of simulated data based on the 25^{th} slice. From left to right: first measure (d_1), second measure (d_2), and third measure (d_3).

It may be that many of the activation differences between control subjects and patients are too small to be detected by Fisher's method. However, the findings of Lazar et al. show Fisher's method to be one of the more highly sensitive combining techniques, capable of finding more areas of task-related brain activation than several comparable methods (Lazar et al., 2002). It is not apparent that a lack of sensitivity is an issue with our method.

4 Conclusions

Ultimately, our methodology identified relatively few regions with significantly different activation between the treatment groups. However, it is not impossible to analyze what we have found. The anterior regions of the 25^{th} through 27^{th} slices are home to anatomical features that coincide with the task being performed. In particular, portions of the middle frontal gyrus can be found in these regions. The middle frontal gyrus plays a major role in target detection (Kirino et al., 2000). Also of importance in the region are the frontal eye fields and the supplementary eye fields. These fields are one of the primary sources of activation during saccade and antisaccade tasks such as those performed in the present study. The frontal eye fields contain neurons that control whether and when eye movement occurs, including the stimuli that result in saccades (Schall, 2002). The supplementary eye fields also contain neurons involved in the saccade process, however they do not appear capable of initiating a saccade on their own (Stuphorn et al., 2010). Our region of interest also contains features that are not directly associated with the antisaccade task in particular. It contains portions of the primary motor cortex, which is responsible for the majority of the body's muscle control. While this may have a tangential impact on the study (blinking, in particular, is one of the cortex's functions), the areas in question are generally associated with movement in the fingers, wrists, and elbows (Penfield and Boldrey, 1937).

Our simulation tests have reassured us of our methodology's ability to detect significant differences in brain activation. However, the incongruity between our results and those achieved by Camchong et al. is curious. Fisher's method has been shown to be more sensitive in its detection of brain activation than random effects models (Lazar et al., 2002), and yet in this case the random effects model used by Camchong et al. has discovered more significant differences in activation. The reasons for this are unclear. However, it is notable that the comparisons performed by Lazar et al. were done on a voxel-by-voxel basis, whereas we grouped regions of voxels together. It is possible that our grouping has masked details that would otherwise have led to the detection of significant groups of voxels. A way to counter this potential masking would be to divide the brain into more than eight regions, but dividing too finely would sacrifice the ability to easily interpret the results, so a balance would have to be struck. It is also possible that the image difference measures are not suitable for such analysis. Alternatively, the findings of Camchong et al. could have contained strong false positive results which were not detected by our methodology.

There are aspects of this study that future research on the subject of schizophrenia may help to shed light on. One factor that might have had an effect on the data was antipsychotic medication. It was being taken by thirteen of the patients at the time of the fMRI scan. No record was kept identifying which of the scans belonged to those taking the medication and which belonged to those who did not, so it could not be accounted for in any analysis of the data. Of course, there are ethical concerns regarding such matters, the chief of which being the privacy of the patients, that may prevent this from being thoroughly investigated. Another factor to consider is the sample size. The sample we have analyzed is relatively small (even without accounting for the fact that we were forced to disregard the scans of two subjects), and smaller sample sizes carry an increased possibility of population misrepresentation. Practical matters are most likely responsible for this, as the organizational

concerns of gathering and scanning individuals can be prohibitively expensive. Still, a larger sample, if possible, could help clarify any ambiguities we may encounter in the analysis of such data.

5 Bibliography

-American Psychiatric Organization (2000). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.). Washington, DC: American Psychiatric Organization.

-Barbato, A. (1999). *Schizophrenia and Public Health*, Geneva: World Health Organization.

-Camchong, J., Dyckman, K., Austin, B., Clementz, B., and McDowell, J. (2008). Common neural circuitry supporting volitional saccades and its disruption in schizophrenia patients and relatives. *Biological Psychiatry*, 64, 1042-1050.

-Christensen, W.F., and Yetkin, Z.F. (2003). A spatio-temporal analytic approach for improved detection of activation in silent fMRI. *Proceedings of the International Society for Magnetic Resonance in Medicine*, 11, 2534.

-Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

-Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

-Fisher, R.A. (1950). *Statistical Methods for Research Workers* (11th ed.). London: Oliver and Boyd.

-Hesterberg, T.C., Moore, D.S., Monaghan, S., Clipson, A., and Epstein, R. (2005). *Introduction to the Practice of Statistics*. New York: W.H. Freeman.

-Huettel, S.A., Song, A.W., and McCarthy, G. (2009). *Functional Magnetic Resonance Imaging* (2nd ed.). Sunderland: Sinauer Associates.

-Kandel E.R., Schwartz, J.H., and Jessell, T.M. (2000). *Principles of Neural Science* (4th ed.). New York: McGraw-Hill.

-Kendler, K.S., McGuire, M., Gruenberg, A.M., O'Hare, A., Spellman, M., and Walsh, D. (1993). The Roscommon Family Study. I. Methods, diagnosis of probands, and risk of schizophrenia in relatives. *Archives of General Psychiatry*, 50, 527-540.

-Kirino, E., Belger, A., Goldman-Rakic, P., and McCarthy, G. (2000). Prefrontal activation evoked by infrequent target and novel stimuli in a visual target detection task: An event-related functional magnetic resonance imaging study. *The Journal of Neuroscience*, 20, 6612-6618.

-Lazar, N.A., Luna, B., Sweeney, J.A., and Eddy, W.F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, 16, 538-550.

-Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453, 869-878.

-Munoz, D.P., and Everling, S. (2004). Look away: The antisaccade task and the voluntary control of eye movement. *Nature Reviews: Neuroscience*, 5, 218-228.

-Ogawa, S., Lee, T.M., Nayak, A.S., and Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14, 68-78.

-Penfield, W., and Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60, 389-443.

-Schall, J.D. (2002). The neural selection and control of saccades by the frontal eye field. *Philosophical Transactions of the Royal Society B*, 357, 1073-1082.

-Smyth, G.K., and Phipson, B. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9, 39.

-Stuphorn, V., Brown, J.W., Schall, J.D. (2010). Role of supplementary eye field in saccade initiation: Executive, not direct, control. *Journal of Neurophysiology*, 103, 801-816.

-Talairach, J., and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. New York: Thieme.

-van der Weken, D., Nachtegael, M., and Kerre, E.E. (2004). Using similarity measures and homogeneity for the comparison of images. *Image and Vision Computing*, 22, 695-702.

-van Os, J., and Kapur, S. (2009). Schizophrenia. *Lancet*, 374, 635-645.

-Wang, W. (1997). New similarity measures on fuzzy sets and on elements. *Fuzzy Set and Systems*, 85, 305-309.

-Welch, B.L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 34, 28-35.